

# Comparing and Aggregating Partially Resolved Trees<sup>†</sup>

Mukul S. Bansal<sup>‡</sup>      Jianrong Dong<sup>‡</sup>      David Fernández-Baca<sup>‡</sup>

## Abstract

We define, analyze, and give efficient algorithms for two kinds of distance measures for rooted and unrooted phylogenies. For rooted trees, our measures are based on the topologies the input trees induce on *triplets*; that is, on three-element subsets of the set of species. For unrooted trees, the measures are based on *quartets* (four-element subsets). Triplet and quartet-based distances provide a robust and fine-grained measure of the similarities between trees. The distinguishing feature of our distance measures relative to traditional quartet and triplet distances is their ability to deal cleanly with the presence of unresolved nodes, also called polytomies. For rooted trees, these are nodes with more than two children; for unrooted trees, they are nodes of degree greater than three.

Our first class of measures are parametric distances, where there is a parameter that weighs the difference between an unresolved triplet/quartet topology and a resolved one. Our second class of measures are based on Hausdorff distance. Each tree is viewed as a set of all possible ways in which the tree could be refined to eliminate unresolved nodes. The distance between the original (unresolved) trees is then taken to be the Hausdorff distance between the associated sets of fully resolved trees, where the distance between trees in the sets is the triplet or quartet distance, as appropriate.

**Keywords.** Aggregation, Hausdorff distance, phylogenetic trees, quartet distance, triplet distance.

## 1 Introduction

Evolutionary trees, also known as phylogenetic trees or phylogenies, represent the evolutionary history of sets of species. Such trees have uniquely labeled leaves, corresponding to the species, and unlabeled internal nodes, representing hypothetical ancestors. The trees can be either rooted, if the evolutionary origin is known, or unrooted, otherwise.

---

<sup>†</sup>An extended abstract of this paper was presented at the 8th Latin American Symposium on Theoretical Informatics, Búzios, Brazil.

<sup>‡</sup>Department of Computer Science, Iowa State University, Ames, IA 50011, USA. Email: {bansal, jdong, fernande}@iastate.edu. The authors were supported in part by National Science Foundation ATOL grants DEB-0334832 and DEB-0829674.

This paper addresses two related questions: (1) How does one measure how close two evolutionary trees are to each other? (2) How does one combine or *aggregate* the phylogenetic information from conflicting trees into a single *consensus tree*? Among the motivations for the first question is the growth of phylogenetic databases, such as TreeBase [28], with the attendant need for sophisticated querying mechanisms and for means to assess the quality of answers to queries. The second question arises from the fact that phylogenetic analyses — e.g., by parsimony [22] — typically produce multiple evolutionary trees (often in the thousands) for the same set of species. Another motivation is the ongoing effort to assemble the tree of life by piecing together phylogenies for subsets of species [17].

We address the above questions by defining appropriate *distance measures* between trees. While several such measures have been proposed before (see below), ours provide a feature that previous ones do not: The ability to deal cleanly with the presence of *unresolved* nodes, also called *polytomies*. For rooted trees these are nodes with more than two children; for unrooted trees, they are nodes of degree greater than three. Polytomies cannot simply be ignored, since they arise naturally in phylogenetic analysis. Furthermore, they must be treated with care: A node may be unresolved because it truly must be so or because there is not enough evidence to break it up into resolved nodes — that is, the polytomies are either “hard” or “soft” [26].

**Our contributions.** We define and analyze two kinds of distance measures for phylogenies. For rooted trees, our measures are based on the topologies the input trees induce on *triplets*; that is, on three-element subsets of the set of species. For unrooted trees, the measures are based on *quartets* (four-element subsets). Our approach is motivated by the observation that triplet and quartet topologies are the basic building blocks of rooted and unrooted trees, in the sense that they are the smallest topological units that completely identify a phylogenetic tree [30]. Triplet and quartet-based distances thus provide a robust and fine-grained measure of the differences and similarities between trees<sup>1</sup>. In contrast with traditional quartet and triplet distances, our two classes of distance measures deal cleanly with the presence of unresolved nodes. Each of them does so in a different way.

The first kind of measures we propose are *parametric distances*: Given a triplet (quartet)  $X$ , we compare the topologies that each of the two input trees induces on  $X$ . If they are identical, the contribution of  $X$  to the distance is zero. If both topologies are fully resolved but different, then the contribution is one. Otherwise, the topology is resolved in one of the trees, but not the other. In this case,  $X$  contributes  $p$  to the distance, where  $p$  is a real number between 0 and 1. Parameter  $p$  allows one to make a smooth transition between hard and soft views of polytomy. At one extreme, if  $p = 1$ , an unresolved topology is viewed as different from a fully resolved one. At the other, when  $p = 0$ , unresolved topologies are viewed as identical to resolved ones. Intermediate values of  $p$  allow one to adjust for the degree of certainty one has about a polytomy.

The second kind of measures proposed here are based on viewing each tree as a set of all possible fully resolved trees that can be obtained from it by refining its unresolved nodes. The distance between two trees is defined as the Hausdorff distance between the corresponding sets<sup>2</sup>,

<sup>1</sup>Biologically-inspired arguments in favor of triplet-based measures can be found in [13].

<sup>2</sup>Informally, two sets  $A$  and  $B$  are at Hausdorff distance  $\tau$  of each other if each element of  $A$  is within distance  $\tau$

where the distance between trees in the sets is the triplet or quartet distance, as appropriate.

After defining our distance measures, we proceed to study their mathematical and algorithmic properties. We obtain exact and asymptotic bounds on expected values of parametric triplet distance and parametric quartet distance. We also study for which values of  $p$ , parametric triplet and quartet distances are metrics, *near-metrics* (in the sense of [19]), or non-metrics.

Aside from the mathematical elegance that metrics and near-metrics bring to tree comparison, there are also algorithmic benefits. We formulate phylogeny aggregation as a *median* problem, in which the objective is to find a consensus tree whose total distance to the given trees is minimized. We do not know whether finding the median tree relative to parametric (triplet or quartet) distance is NP-hard, but conjecture that it is. This is suggested by the NP-completeness of the *maximum triplet compatibility problem*<sup>3</sup> [9]. However, by the results mentioned above and well-known facts about the median problem [36], there are simple constant-factor approximation algorithms for the aggregation of rooted and unrooted trees relative to parametric distance: Simply return the input tree with minimum distance to the remaining input trees. We show that there are values of  $p$  for which parametric distance is a metric, but the median tree may not be fully resolved even if all the input trees are. However, beyond a threshold, the median tree is guaranteed to be fully resolved if the input trees are fully resolved.

A natural problem is whether Hausdorff triplet (quartet) distance between two trees can be computed in polynomial time. We suspect that computing Hausdorff triplet (quartet) distance is NP-hard. However, even if this were so, we show that one can partially circumvent the issue by proving that, under a certain density assumption, Hausdorff distance is within a constant factor of parametric distance — that is, the measures are *equivalent* in the sense of [19].

Finally, we present a  $O(n^2)$ -time algorithm to compute parametric triplet distance and a  $O(n^2)$  2-approximate algorithm for parametric quartet distance. To our knowledge, there was no previous algorithm for computing the parametric triplet distance between two rooted trees, other than by enumerating all  $\Theta(n^3)$  triplets. Two algorithms exist that can be directly applied to compute the parametric quartet distance (see also [11]). One runs in time  $O(n^2 \min\{d_1, d_2\})$ , where, for  $i \in \{1, 2\}$ ,  $d_i$  is the maximum degree of a node in  $T_i$  [12]; the other takes  $O(d^9 n \log n)$  time, where  $d$  is the maximum degree of a node in  $T_1$  and  $T_2$  [34].<sup>4</sup> Our faster  $O(n^2)$  algorithm offers a 2-approximate solution when an exact value of the parametric quartet distance is not required. Additionally, our algorithm gives the exact answer when  $p = \frac{1}{2}$ .

**Related work.** Several other measures for comparing trees have been proposed; we mention a few. A popular class of distances are those based on symmetric distance between sets of *clusters* (that is, on sets of species that descend from the same internal node in a rooted tree) or of *splits* (partitions of the set of species induced by the removal of an edge in an unrooted tree); the latter is the well-known Robinson-Foulds (RF) distance [29]. It is not hard to show that two rooted

---

of  $B$  and vice-versa. For a formal definition, see Section 3.

<sup>3</sup>The input to this problem consists of a set of trees, each of which has three leaves; the leaf sets of these trees may not be identical. The question is to find the largest subset of these triplet trees such that all of the trees are consistent with a single tree  $T$  whose leaf set is the union of the leaves of the input triplet trees.

<sup>4</sup>Note that the presence of unresolved nodes seems to complicate distance computation. Indeed, the quartet distance between a pair of *fully resolved* unrooted trees can be obtained in  $O(n \log n)$  time [8].

(unrooted) trees can share many triplet (quartet) topologies but not share a single cluster (split). Cluster- and split-based measures are also coarser than triplet and quartet distances.

One can also measure the distance between two trees by counting the number of *branch-swapping* operations — e.g., nearest-neighbor interchange or subtree pruning and regrafting operations [22] — needed to convert one of the trees into the other [3]. However, the associated measures can be hard to compute, and they fail to distinguish between operations that affect many species and those that affect only a few. An alternative to distance measures are *similarity* methods such as maximum agreement subtree (MAST) approach [23]. While there are efficient algorithms for computing the MAST [21], the measure is coarser than triplet-based distances.

There is an extensive literature on consensus methods for phylogenetic trees. A non-exhaustive list of methods based on splits or clusters includes strict consensus trees [27], majority-rule trees [4], and the Adams consensus [1]. In *local consensus* methods, the goal is to find a consensus tree that satisfies a given set of constraints on the topology of each triplet [24]. For a thorough survey of these methods, their properties and interrelationships, see [10].

The fact that consensus methods tend to produce unresolved trees, with an attendant loss of information, has been observed before. An alternative approach is to provide multiple consensus trees, instead of a single one. The idea, developed more fully in [35], is to cluster the input trees using some distance measure into groups, each of which is represented by a single consensus tree, in such a way as to minimize some measure of information loss. Our distance measures can be used within this framework, where their fine-grained nature could conceivably offer advantages over other techniques.

In addition to consensus methods, there are techniques that take as input sets of quartet trees or triplet trees and try to find large compatible subsets or subsets whose removal results in a compatible set [6, 31]. These problems are related to the *supertree problem*, in which a set of input trees that may not all share the same species is given and the problem is to find a single tree that exhibits as much as possible of the evolutionary relationships among the input trees [7]. Thus, the consensus problem for trees is a special case of the supertree problem.

The consensus problem on trees exhibits parallels with the *rank aggregation problem*, a problem with a rich history and which has recently found applications to Internet search [2, 5, 14, 16, 25, 18, 19]. Here, we are given a collection of rankings (that is, permutations) of  $n$  objects, and the goal is to find a ranking of minimum total distance to the input rankings. A distance between rankings of particular interest is *Kendall's tau*, defined as the number of pairwise disagreements between the two rankings. Like triplet and quartet distances, Kendall's tau is based on elementary ordering relationships. Rank aggregation under Kendall's tau was shown to be NP-complete even for four lists by Dwork et al. [18].

A permutation is the analog of a fully resolved tree, since every pairwise relationship between elements is given. The analog to a partially-resolved tree is a *partial ranking*, in which the elements are grouped into an ordered list of *buckets*, such that elements in different buckets have known ordering relationships, but elements within a bucket are not ranked [19]. Our definitions of parametric distance and Hausdorff distance are inspired by Fagin et al.'s *Kendall tau with parameter  $p$*  and their Hausdorff version of Kendall's tau, respectively [19]. We note, however, that aggregating partial rankings seems computationally easier than the consensus problem on trees.

For example, while the Hausdorff version of Kendall’s tau has a simple and easily-computable expression [14, 19], it is unclear whether the Hausdorff triplet or quartet distances are polynomially-computable for trees.

**Organization of the paper.** Section 2 reviews basic notions in phylogenetics and distances. Our distance measures and the consensus problem are formally defined in Section 3. The expected values of the distance measures are studied in Section 4. The basic properties of parametric distance are proved in Section 5. Section 6 studies the connection between Hausdorff and parametric distances. Section 7 gives efficient algorithms for computing parametric triplet distance. A 2-approximation algorithm for parametric quartet distance is given in Section 8.

## 2 Preliminaries

**Phylogenies.** By and large, we follow standard terminology (i.e., similar to [9] and [30]). We write  $[N]$  to denote the set  $\{1, 2, \dots, N\}$ , where  $N$  is a positive integer.

Let  $T$  be a rooted or unrooted tree. We write  $\mathcal{V}(T)$ ,  $\mathcal{E}(T)$ , and  $\mathcal{L}(T)$  to denote, respectively, the node set, edge set, and leaf set of  $T$ . A *taxon* (plural *taxa*) is some basic unit of classification; e.g., a species. Let  $S$  be a set of taxa. A *phylogenetic tree* or *phylogeny* for  $S$  is a tree  $T$  such that  $\mathcal{L}(T) = S$ . Furthermore, if  $T$  is rooted, we require that every internal node have at least two children; if  $T$  is unrooted, every internal node is required to have degree at least three. We write  $RP(n)$  to denote the set of all rooted phylogenetic trees over  $S = [n]$  and  $P(n)$  to denote the set of all unrooted phylogenetic trees over  $S = [n]$ .

An internal node in a *rooted* phylogeny is *resolved* if it has exactly two children; otherwise it is *unresolved*. Similarly, an internal node in an *unrooted* phylogeny is *resolved* if it has degree three, and *unresolved* otherwise. Unresolved nodes in rooted and unrooted trees are also referred to as *polytomies* or *multifurcations*. A phylogeny (rooted or unrooted) is *fully resolved* if all its internal nodes are resolved. A *fan* is a completely unresolved phylogeny; i.e., it contains a single internal node, to which all leaves are connected (if the phylogeny is rooted, this internal node is the root).

A *contraction* of a phylogeny  $T$  is obtained by deleting an internal edge and identifying its endpoints. A phylogeny  $T_2$  is a *refinement* of phylogeny  $T_1$ , denoted  $T_1 \preceq T_2$ , if and only if  $T_1$  can be obtained from  $T_2$  through 0 or more contractions. Tree  $T_2$  is a *full refinement* of  $T_1$  if  $T_1 \preceq T_2$  and  $T_2$  is fully resolved. We write  $\mathcal{F}(T)$  to denote the set of all full refinements of  $T$ .

Let  $X$  be a subset of  $\mathcal{L}(T)$  and let  $T[X]$  denote the minimal subtree of  $T$  having  $X$  as its leaf set. The *restriction* of  $T$  to  $X$ , denoted  $T|X$ , is the phylogeny for  $X$  defined as follows. If  $T$  is unrooted, then  $T|X$  is the tree obtained from  $T[X]$  by suppressing all degree-two nodes. If  $T$  is rooted,  $T|X$  is obtained from  $T[X]$  by suppressing all degree-two nodes except for the root.

A *triplet* is a three-element subset of  $S$ . A *triplet tree* is a rooted phylogeny whose leaf set is a triplet. The triplet tree with leaf set  $\{a, b, c\}$  is denoted by  $a|bc$  if the path from  $b$  to  $c$  does not intersect the path from  $a$  to the root. A *quartet* is a four-element subset of  $S$  and a *quartet tree* is an unrooted phylogeny whose leaf set is a quartet. The quartet tree with leaf set  $\{a, b, c, d\}$  is denoted by  $ab|cd$  if the path from  $a$  to  $b$  does not intersect the path from  $c$  to  $d$ . A triplet (quartet)

$X$  is said to be *resolved* in a phylogenetic tree  $T$  over  $S$  if  $T|X$  is fully resolved; otherwise,  $X$  is *unresolved*.

Finally, we introduce notation for certain useful subtrees of a tree  $T$ . Suppose  $T$  is rooted and  $v$  is a node in  $T$ . Then,  $T(v)$  denotes the subtree of  $T$  rooted at  $v$ . Suppose  $T$  is unrooted and  $\{u, v\}$  is an edge in  $T$ . Removal of edge  $\{u, v\}$  splits the tree  $T$  into two subtrees. We denote the subtree that contains node  $u$  by  $T(u, v)$ , and the subtree that contains  $v$  by  $T(v, u)$ .

**Distance measures, metrics, and near-metrics.** A *distance measure* on a set  $D$  is a binary function  $d$  on  $D$  satisfying the following three conditions: (i)  $d(x, y) \geq 0$  for all  $x, y \in D$ ; (ii)  $d(x, y) = d(y, x)$  for all  $x, y \in D$ ; and (iii)  $d(x, y) = 0$  if and only if  $x = y$ . Function  $d$  is a *metric* if, in addition to being a distance measure, it satisfies the triangle inequality; i.e.,  $d(x, z) \leq d(x, y) + d(y, z)$  for all  $x, y, z \in D$ . Distance measure  $d$  is a *near-metric* if there is a constant  $c$ , independent of the size of  $D$ , such that  $d$  satisfies the *relaxed polygonal inequality*:  $d(x, z) \leq c(d(x, x_1) + d(x_1, x_2) + \dots + d(x_{n-1}, z))$  for all  $n > 1$  and  $x, z, x_1, \dots, x_{n-1} \in D$  [19]. Two distance measures  $d$  and  $d'$  with domain  $D$  are *equivalent* if there are constants  $c_1, c_2 > 0$  such that  $c_1 d'(x, y) \leq d(x, y) \leq c_2 d'(x, y)$  for every pair  $x, y \in D$  [19].

### 3 Distance measures for phylogenies

Here we define the distance measures for rooted and unrooted trees to be studied in the rest of the paper. We use essentially the same notation for the rooted tree measures as for the unrooted tree measures. We do so because the concepts for each case are close analogs of those for the other, the key difference being the use of triplets in one setting (rooted trees) and of quartets in the other (unrooted trees). It will be easy to distinguish between the two settings by simply specifying the context in which the measures are being applied. Our notation has the benefits of reducing repetitiveness and of allowing us to avoid excessive use of subscripts and superscripts.

Let  $T_1$  and  $T_2$  be any two rooted (respectively, unrooted) phylogenies over taxon set  $[n]$ . Define the following five sets of triplets (quartets) over  $[n]$ .

$\mathcal{S}(T_1, T_2)$ : The set of all triplets (quartets)  $X$  such that  $T_1|X$  and  $T_2|X$  are fully resolved, and  $T_1|X = T_2|X$ .

$\mathcal{D}(T_1, T_2)$ : The set of all triplets (quartets)  $X$  such that  $T_1|X$  and  $T_2|X$  are fully resolved, and  $T_1|X \neq T_2|X$ .

$\mathcal{R}_1(T_1, T_2)$ : The set of all triplets (quartets)  $X$  such that  $T_1|X$  is fully resolved, but  $T_2|X$  is not.

$\mathcal{R}_2(T_1, T_2)$ : The set of all triplets (quartets)  $X$  such that  $T_2|X$  is fully resolved, but  $T_1|X$  is not.

$\mathcal{U}(T_1, T_2)$ : The set of all triplets (quartets)  $X$  such that  $T_1|X$  and  $T_2|X$  are unresolved.

Let  $p$  be a real number in the interval  $[0, 1]$ . The *parametric triplet (quartet) distance* between  $T_1$  and  $T_2$  is defined as<sup>5</sup>

$$d^{(p)}(T_1, T_2) = |\mathcal{D}(T_1, T_2)| + p(|\mathcal{R}_1(T_1, T_2)| + |\mathcal{R}_2(T_1, T_2)|). \quad (1)$$

When the domain of  $d^{(p)}$  is restricted to fully resolved trees, and thus  $\mathcal{R}_1(T_1, T_2) = \mathcal{R}_2(T_1, T_2) = \mathcal{U}(T_1, T_2) = \emptyset$ , we refer to it simply as the *triplet (quartet) distance*.

Parameter  $p$  allows one to make a smooth transition from soft to hard views of polytomy: When  $p = 0$ , resolved triplets (quartets) are treated as equal to unresolved ones, while when  $p = 1$ , they are treated as being completely different. Choosing intermediate values of  $p$  allows one to adjust for the amount of evidence required to resolve a polytomy<sup>6</sup>.

An alternative distance measure (inspired by References [19, 14]), is the *Hausdorff distance*, defined as follows. Let  $d$  be a metric over fully resolved trees. Metric  $d$  is extended to partially resolved trees as follows.

$$d_{\text{Haus}}(T_1, T_2) = \max \left\{ \max_{t_1 \in \mathcal{F}(T_1)} \min_{t_2 \in \mathcal{F}(T_2)} d(t_1, t_2), \max_{t_2 \in \mathcal{F}(T_2)} \min_{t_1 \in \mathcal{F}(T_1)} d(t_1, t_2) \right\} \quad (2)$$

When  $d$  is the triplet (quartet) distance,  $d_{\text{Haus}}$  is called the *Hausdorff triplet (quartet) distance*.

Definition (2) requires some explanation. The quantity  $\min_{t_2 \in \mathcal{F}(T_2)} d(t_1, t_2)$  is the distance between  $t_1$  and the set of full refinements of  $T_2$ . Hence,

$$\max_{t_1 \in \mathcal{F}(T_1)} \min_{t_2 \in \mathcal{F}(T_2)} d(t_1, t_2)$$

is the maximum distance between a full refinement of  $T_1$  and the set of full refinements of  $T_2$ . Similarly,

$$\max_{t_2 \in \mathcal{F}(T_2)} \min_{t_1 \in \mathcal{F}(T_1)} d(t_1, t_2)$$

is the maximum distance between a full refinement of  $T_2$  and the set of full refinements of  $T_1$ . Therefore,  $T_1$  and  $T_2$  are at Hausdorff distance  $r$  of each other if every full refinement of  $T_1$  is within distance  $r$  of a full refinement of  $T_2$  and vice-versa.

**Aggregating phylogenies.** Let  $k$  be a positive integer and  $S$  be a set of taxa. A *profile of length  $k$*  (or simply a *profile*, when  $k$  is understood from the context) is a mapping  $\mathcal{P}$  that assigns each  $i \in [k]$  a phylogenetic tree  $\mathcal{P}(i)$  over  $S$ . We refer to these trees as *input trees*. A *consensus rule* is a function that maps a profile  $\mathcal{P}$  to some phylogenetic tree  $T$  over  $S$  called a *consensus tree*.

Let  $d$  be a distance measure whose domain is the set of phylogenies over  $S$ . We extend  $d$  to define a distance measure from profiles to phylogenies as  $d(T, \mathcal{P}) = \sum_{i=1}^k d(T, \mathcal{P}(i))$ . A consensus rule is a *median rule* for  $d$  if for every profile  $\mathcal{P}$  it returns a phylogeny  $T^*$  of minimum distance to  $\mathcal{P}$ ; such a  $T^*$  is called a *median*. The problem of finding a median for a profile with respect to a distance measure  $d$  is referred to as the *median problem* (relative  $d$ ), or as the *aggregation problem*.

<sup>5</sup>Note that the sets  $\mathcal{S}(T_1, T_2)$  and  $\mathcal{U}(T_1, T_2)$  are not used in the definition of  $d^{(p)}$ , but are needed for other purposes.

<sup>6</sup>We note that parametric triplet/quartet distance is a *profile-based metric*, in the sense of [19]. However, the use of the word “profile” in [19] is quite different from our use of the term.

## 4 Expected parametric triplet and quartet distances

We now consider the expected value of parametric triplet and quartet distances. Let  $u(n)$  and  $r(n)$  denote the probabilities that a given quartet is, respectively, unresolved or resolved in an unrooted phylogeny chosen uniformly at random from  $P(n)$ ; thus,  $u(n) = 1 - r(n)$ . The following are the two main results of this section.

**Theorem 4.1.** *Let  $T_1$  and  $T_2$  be two unrooted phylogenies chosen uniformly at random with replacement from  $P(n)$ . Then,*

$$E(d^{(p)}(T_1, T_2)) = \binom{n}{4} \cdot \left( \frac{2}{3} \cdot r(n)^2 + 2 \cdot p \cdot r(n) \cdot u(n) \right). \quad (3)$$

**Theorem 4.2.** *Let  $T_1$  and  $T_2$  be two rooted phylogenies chosen uniformly at random with replacement from  $RP(n)$ . Then,*

$$E(d^{(p)}(T_1, T_2)) = \binom{n}{3} \cdot \left( \frac{2}{3} \cdot r(n+1)^2 + 2 \cdot p \cdot r(n+1) \cdot u(n+1) \right). \quad (4)$$

It is known [33, 32] that

$$u(n) \sim \sqrt{\frac{\pi(2 \ln 2 - 1)}{4n}}. \quad (5)$$

Together with Theorems 4.1 and 4.2, this implies that  $E(d^{(p)}(T_1, T_2))$  is asymptotically  $\frac{2}{3} \cdot \binom{n}{4}$  for unrooted trees and  $\frac{2}{3} \cdot \binom{n}{3}$  for rooted trees.

The proof of Theorem 4.1 follows directly from the work of Day [15]; hence, it is omitted (however, we should note that the proof is similar to that of Lemma 4.1 below). In the remainder of this section, we give a proof of Theorem 4.2.

We need some notation. Let  $u'(n)$  and  $r'(n)$  denote the probabilities that a given triplet is, respectively, unresolved or resolved in an rooted phylogeny chosen at random from  $RP(n)$ .

**Lemma 4.1.** *Let  $T_1$  and  $T_2$  be two rooted phylogenies chosen uniformly at random with replacement from  $RP(n)$ . Then,*

$$E(d^{(p)}(T_1, T_2)) = \binom{n}{3} \cdot \left( \frac{2}{3} \cdot r'(n)^2 + 2 \cdot p \cdot r'(n) \cdot u'(n) \right). \quad (6)$$

*Proof.* By the definition of  $d^{(p)}$  and the linearity of expectation, it suffices to establish the equalities below.

$$E(\mathcal{D}(T_1, T_2)) = \binom{n}{3} \cdot \frac{2}{3} \cdot r'(n)^2 \quad (7)$$

$$E(\mathcal{R}_1(T_1, T_2)) = E(\mathcal{R}_2(T_1, T_2)) = \binom{n}{3} \cdot r'(n) \cdot u'(n) \quad (8)$$

To establish Equation (7), consider a triplet  $X$ . The probability that  $X$  is resolved in  $T_1$  (or  $T_2$ ) is  $r'(n)$ . Thus, the probability that  $X$  is resolved in both  $T_1$  and  $T_2$  is  $r'(n)^2$ . There are exactly



three different ways in which any given triplet can be resolved. Hence, if  $\alpha$  is resolved in both  $T_1$  and  $T_2$ , the probability that it is resolved differently in both trees is  $\frac{2}{3}$ . Thus, the probability of a pre-given triplet being resolved in both  $T_1$  and  $T_2$ , but with different types in each, is  $\frac{2}{3}r'(n)^2$ . By the linearity of expectation and since the total number of triplets from  $\mathcal{L}(T_1)$  (and  $\mathcal{L}(T_2)$ ) is  $\binom{n}{3}$ ,  $E(\mathcal{D}(T_1, T_2)) = \binom{n}{3} \cdot \frac{2}{3}r'(n)^2$ .

To establish Equation (8), we only need to study  $E(\mathcal{R}_1(T_1, T_2))$ ; the expression for  $E(\mathcal{R}_2(T_1, T_2))$  follows by symmetry. Consider a triplet  $X$ . The probability that  $X$  is unresolved in  $T_1$  is  $u'(n)$  and the probability that  $X$  is resolved in  $T_2$  is  $r'(n)$ . The expression for  $E(\mathcal{R}_1(T_1, T_2))$  now follows by linearity of expectation.  $\square$

Let us define the function  $\text{ADD-LEAF} : RP(n) \rightarrow P(n+1)$  as follows. Given a rooted tree  $T \in RP(n)$ ,  $\text{ADD-LEAF}(T)$  is the unrooted tree constructed from  $T$  by (1) adding a leaf node labeled  $n+1$  to  $T$  by adjoining it to the root node of  $T$  and (2) unrooting the resulting tree. The next two lemmas are well known (for proofs, see [33, 22] and [30, p. 20], respectively).

**Lemma 4.2.** *For all  $n \geq 1$ ,  $|RP(n)| = |P(n+1)|$ .*

**Lemma 4.3.** *Function  $\text{ADD-LEAF}$  is a bijection from the set  $RP(n)$  to the set  $P(n+1)$ .*

For any triplet  $X$  over  $[n]$ , we define two functions  $g_X : RP(n) \rightarrow \{0, 1\}$  and  $f_X : P(n+1) \rightarrow \{0, 1\}$  as follows:

$$g_X(T) = \begin{cases} 1 & \text{if triplet } X \text{ is resolved in tree } T \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$f_X(T) = \begin{cases} 1 & \text{if quartet } X \cup \{n+1\} \text{ is resolved in tree } T \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

We have the following result.

**Lemma 4.4.** *Let  $X$  be any triplet over  $[n]$ . Consider a tree  $T \in RP(n)$ , and let  $T' = \text{ADD-LEAF}(T)$ . Then,  $f_X(T') = g_X(T)$ .*

*Proof.* Follows from the observation that triplet  $X$  is resolved in  $T$  if and only if quartet  $X \cup \{n+1\}$  is resolved in  $T'$ .  $\square$

**Lemma 4.5.** *For all  $n \geq 1$ ,  $r'(n) = r(n+1)$  and  $u'(n) = u(n+1)$ .*

*Proof.* Let  $X$  be any triplet over  $[n]$ . By definition,  $r(n+1)$  is the probability of any given quartet being resolved in a random unrooted tree in  $P(n)$ . In particular,  $r(n+1)$  is the probability that

quartet  $X \cup \{n+1\}$  is resolved in a random unrooted tree. Now,

$$\begin{aligned}
r(n+1) &= \sum_{T \in P(n+1)} \frac{f_X(T)}{|P(n+1)|} \\
&= \sum_{T \in P(n+1)} \frac{f_X(T)}{|RP(n)|} \\
&= \sum_{T' \in RP(n)} \frac{g_X(T')}{|RP(n)|} \\
&= r'(n),
\end{aligned}$$

where the first and last equalities follow from the definitions of  $r(n+1)$  and  $r(n)$ , respectively, the second equality follows from Lemma 4.2, and the third follows from Lemma 4.3 and Lemma 4.4.

Since  $u'(n) = 1 - r'(n)$  and  $u(n+1) = 1 - r(n+1)$ , it follows that  $u'(n) = u(n+1)$ .  $\square$

*Proof of Theorem 4.2.* Simply substitute the expressions for  $r'(n)$  and  $u'(n)$  given in Lemma 4.5 into the expression for  $E(d^{(p)}(T_1, T_2))$  given in Lemma 4.1.  $\square$

## 5 Properties of parametric distance

In what follows, unless mentioned explicitly, whenever we refer to parametric distance, we mean both its triplet and quartet varieties. We begin with a useful observation.

**Proposition 5.1.** *For every  $p, q$  such that  $p, q \in (0, 1]$ ,  $d^{(p)}$  and  $d^{(q)}$  are equivalent.*

*Proof.* Let  $T_1$  and  $T_2$  be two rooted (unrooted) trees. Let  $M$  be the number of triplets (quartets) resolved differently in  $T_1$  and let  $N$  be the number of triplets (quartets) resolved only in one of  $T_1$  and  $T_2$ . Then,  $d^{(p)}(T_1, T_2) = M + pN$ , and  $d^{(q)}(T_1, T_2) = M + qN$ . Without loss of generality, let  $p \geq q$ . Now, if  $c_1 = q/p$ , then we have  $c_1 d^{(q)}(T_1, T_2) = qM/p + q^2 N/p \leq M + pN = d^{(p)}(T_1, T_2)$ . Similarly, if  $c_2 = p/q$ , then we have  $c_2 d^{(q)}(T_1, T_2) = pM/q + pN \geq M + pN = d^{(p)}(T_1, T_2)$ . Thus,  $c_1 d^{(q)}(T_1, T_2) \leq d^{(p)}(T_1, T_2) \leq c_2 d^{(q)}(T_1, T_2)$ , and, consequently,  $d^{(p)}$  and  $d^{(q)}$  are equivalent.  $\square$

The next result precisely characterizes the ranges of  $p$  for which  $d^{(p)}$  is a metric or near-metric:

**Theorem 5.1.**

- (i) For  $p = 0$ ,  $d^{(p)}$  is not a distance measure.
- (ii) For  $p \in (0, 1/2)$ ,  $d^{(p)}$  is a distance measure, but not a metric.
- (iii) For  $p \in [1/2, 1]$ ,  $d^{(p)}$  is a metric.
- (iv) For  $p \in (0, 1/2)$ ,  $d^{(p)}$  is a near-metric.

*Proof.* Our proof is analogous to the proof of the corresponding result for partial rankings given by Fagin et al. [19]. For the sake of completeness, we prove this result formally. For concreteness, we state our arguments in terms of rooted trees and triplets. The extension to unrooted trees and quartets is direct.

For the proof of (i) and (ii), we use the same three triplet trees,  $t_1 = ab|c$ ,  $t_2 = abc$  (i.e., a completely unresolved tree), and  $t_3 = ac|b$ . To prove (i), we note that  $d^{(0)}(t_1, t_2) = 0$ , even though  $t_1 \neq t_2$ . Thus  $d^{(0)}$  is not a distance measure. Observe also that  $d^{(0)}$  violates the triangle inequality, since  $d^{(0)}(t_1, t_2) + d^{(0)}(t_2, t_3) = 2p = 0 < 1 = d^{(0)}(t_1, t_3)$ .

To prove (ii), observe that  $d^{(p)}(t_1, t_2) = d^{(p)}(t_2, t_3) = p$ , and  $d^{(p)}(t_1, t_3) = 1$ . Thus,  $d^{(p)}(t_1, t_3) = 1 > 2p = d^{(p)}(t_1, t_2) + d^{(p)}(t_2, t_3)$ , violating the triangle inequality. Thus,  $d^{(p)}$  is not a metric in this case. On the other hand, it is straightforward to verify that for any  $p \in (0, 1/2)$  — as well, indeed, as for any  $p \in [1/2, 1]$  — and any trees  $T_1$  and  $T_2$ , we have  $d^{(p)}(T_1, T_2) \geq 0$ ,  $d^{(p)}(T_1, T_2) = d^{(p)}(T_2, T_1)$ , and  $d^{(p)}(T_1, T_2) = 0$  if and only if  $T_1 = T_2$ . Thus,  $d^{(p)}$  is a distance measure in this case.

We now prove (iii). As mentioned in the proof of part (ii),  $d^{(p)}$  is a distance measure for  $p \in [1/2, 1]$ . To complete the proof, we show that the triangle inequality holds; i.e.,  $d^{(p)}(T_1, T_3) \leq d^{(p)}(T_1, T_2) + d^{(p)}(T_2, T_3)$  for any three trees  $T_1, T_2, T_3$ . Note that for any  $i, j \in \{1, 2, 3\}$ , we can express  $d^{(p)}(T_i, T_j)$  as

$$d^{(p)}(T_i, T_j) = \sum_{\{a,b,c\} \subseteq [n]} d^{(p)}(T_i|_{\{a,b,c\}}, T_j|_{\{a,b,c\}}).$$

That is, the distance between  $T_i$  and  $T_j$  can be expressed as the sum of parametric distances between all possible triplet trees induced by  $T_i$  and  $T_j$ . For any  $\{a, b, c\} \subseteq [n]$ , and each  $i \in \{1, 2, 3\}$ , let  $t_i = T_i|_{\{a,b,c\}}$ . To complete the proof of (iii), it suffices to show that  $d^{(p)}(t_1, t_3) \leq d^{(p)}(t_1, t_2) + d^{(p)}(t_2, t_3)$ . If  $t_1 = t_3$ , then  $d^{(p)}(t_1, t_3) = 0 \leq d^{(p)}(t_1, t_2) + d^{(p)}(t_2, t_3)$ , since distances are nonnegative. If  $t_1 \neq t_3$ , then  $d^{(p)}(t_1, t_3) \leq 1$ , while  $d^{(p)}(t_1, t_2) + d^{(p)}(t_2, t_3) \geq 2p$ . Thus,  $d^{(p)}(t_1, t_3) \leq d^{(p)}(t_1, t_2) + d^{(p)}(t_2, t_3)$  if  $p \in [1/2, 1]$ .

Finally, we prove (iv). By Proposition 5.1, for every  $p \in (0, 1/2)$ ,  $d^{(p)}$  is equivalent to  $d^{(1/2)}$ , which, by part (iii), is a metric. The claim now follows from a result by Fagin et al. [20] that implies that a distance measure is a near metric if and only if it is equivalent to a metric.  $\square$

Part (iii) of Theorem 5.1 leads directly to approximation algorithms: Let  $\mathcal{P}$  be a profile, let  $T^*$  be the median tree for  $\mathcal{P}$ , and let  $T = \mathcal{P}(\ell)$ , where  $\ell = \arg \min_i d(\mathcal{P}(i), \mathcal{P})$ . Then, by a standard approximation bound argument (e.g., like those found in [36]), we have that  $d(T, \mathcal{P}) \leq 2d(T^*, \mathcal{P})$ . Part (iv) indicates that the measure degrades nicely, since, along with the 2-approximation algorithm for  $p \in [1/2, 1]$  implied by (iii), it leads to constant factor approximation algorithms for  $p \in (0, 1/2)$  (an analogous observation for aggregation of partial rankings is made in [19]).

The next result establishes a threshold for  $p$  beyond which a collection of fully resolved trees give enough evidence to produce a fully resolved tree, despite the disagreements among them.

**Theorem 5.2.** *Let  $\mathcal{P}$  be a profile of length  $k$ , such that for all  $i \in [k]$ , tree  $\mathcal{P}(i)$  is fully resolved. Then, if  $p \geq 2/3$ , there exists median tree  $T$  for  $\mathcal{P}$  relative to  $d^{(p)}$  such that  $T$  is fully resolved.*

It is interesting to compare Theorem 5.2 with analogous results for partial rankings. Consider the variation of Kendall's tau for partial rankings in which a pair of items that is ordered in one ranking but in the same bucket in the other contributes  $p$  to the distance, where  $p \in [0, 1]$ . This distance measure is a metric when  $p \geq 1/2$  [19]. Furthermore, if  $p \geq 1/2$  the median ranking relative to this distance (that is, the one that minimizes the total distance to the input rankings) is a full ranking if the input consists of full rankings [5]. In contrast, Proposition 5.1 and Theorem 5.2 show that, in the range  $p \in [1/2, 2/3)$ , parametric triplet or quartet distance are metrics, but the median tree is not guaranteed to be fully resolved even if the input trees are. The intuitive reason is that for rankings, there are only two possible outcomes for a comparison between two elements, but there are three ways in which a triplet or quartet may be resolved. This opens up a potentially useful range of values for  $p$  wherein parametric triplet/quartet distance is a metric, but where one can adjust for the degree of evidence (or confidence) needed to resolve a node.

Our proof of Theorem 5.2 relies on two lemmas, which make use of the two procedures below.

**PULL-OUT**( $T, u$ ): The arguments are a rooted phylogenetic tree  $T$  and a non-root node  $u$  in  $T$ , whose parent, denoted by  $v$ , has 3 or more children. The procedure returns a new tree  $T'$  obtained from  $T$  as follows. Split  $v$  into two nodes  $v'$  and  $v''$  such that the parent of  $v'$  equals the parent of  $v$ , the children of  $v'$  are  $u$  and  $v''$ , and the children of  $v''$  are all the children of  $v$  except for  $u$ .

**PULL-2-OUT**( $T, u_1, u_2$ ): The arguments are an unrooted phylogenetic tree  $T$  and two nodes  $u_1, u_2$  sharing the same neighbor  $v$  whose degree is at least 4 in  $T$ . The procedure returns a new tree  $T'$  obtained from  $T$  as follows. Split  $v$  into two nodes  $v'$  and  $v''$  such that the neighbors of  $v'$  are  $v''$ ,  $u_1$ , and  $u_2$ , the neighbors of  $v''$  are  $v'$  and the neighbors of  $v$  except for  $u_1$  and  $u_2$ .

In what follows, we write  $T_i$  to denote  $\mathcal{P}(i)$ , the  $i$ -th tree in profile  $\mathcal{P}$ , for  $i \in [k]$ . We need to introduce separate but analogous concepts for rooted and unrooted trees.

Suppose  $T$  is a rooted phylogenetic tree and let  $v$  be any node in  $T$  with at least 3 children, denoted  $u_1, u_2, \dots, u_d$ . For  $q \in [d]$ , let  $T^{(q)} = \text{PULL-OUT}(T, u_q)$  and let  $L_q$  denote the set of triplets  $X$  such that  $T|X$  is not fully resolved but  $T^{(q)}|X$  is fully resolved. Define the following two quantities.

$$f_q = \sum_{X \in L_q} |\{i \in [k] : T_i|X \text{ agrees with } T^{(q)}|X\}| \quad (11)$$

$$a_q = \sum_{X \in L_q} |\{i \in [k] : T_i|X \text{ disagrees with } T^{(q)}|X\}|. \quad (12)$$

Informally,  $f_q$  and  $a_q$  are the number of *votes* cast by the trees in profile  $\mathcal{P}$  for and against the way the triplets in  $L_q$  are resolved in  $T^{(q)}$ . Indeed, note that, by assumption, every tree in profile  $\mathcal{P}$  is fully resolved. Thus, for each triplet  $X = \{x, y, z\}$  and every  $i \in [k]$ ,  $T_i|X$  must agree with exactly one of  $x|yz$ ,  $y|xz$ , or  $z|xy$ . Thus, there are  $k$  votes associated with each triplet  $X$ , some for, some against.

Now suppose  $T$  is an unrooted phylogenetic tree. Let  $v$  be any node in phylogeny  $T$  and let  $u_1, u_2, \dots, u_d$  be the neighbors of  $v$ . For  $q, r \in [d]$ , let  $T^{(qr)} = \text{PULL-2-OUT}(T, u_q, u_r)$  and let  $L_{qr}$  denote the set of quartets  $X$  such that  $T|X$  is not fully resolved but  $T^{(qr)}|X$  is fully resolved. Define the following two quantities.

$$f_{qr} = \sum_{X \in L_{qr}} |\{i \in [k] : T_i|X \text{ agrees with } T^{(qr)}|X\}| \quad (13)$$

$$a_{qr} = \sum_{X \in L_{qr}} |\{i \in [k] : T_i|X \text{ disagrees with } T^{(qr)}|X\}|. \quad (14)$$

We have the following result.

**Lemma 5.1.** *For the rooted case, there exists an index  $q \in [d]$  such that  $f_q \geq a_q/2$ . For the unrooted case, there exists two indices  $q, r \in [d]$  such that  $f_{qr} \geq a_{qr}/2$ .*

*Proof.* For the rooted case, let  $L = \bigcup_{q=1}^d L_q$ . Thus,  $L$  consists of those triplets that are unresolved in  $T$ , but resolved in  $T^{(q)}$ , for some  $q \in [d]$ . Equivalently,  $L$  consists of those triplets whose elements are leaves from three different subtrees of  $v$ .

Let  $X = \{x, y, z\}$  be a triplet in  $L$ . Assume that  $x \in \mathcal{L}(T(u_q))$ ,  $y \in \mathcal{L}(T(u_r))$ , and  $z \in \mathcal{L}(T(u_s))$ , where  $q, r, s$  must be distinct indices in  $[d]$ . Then,  $X$  is in  $L_q, L_r$ , and  $L_s$ .

Consider any  $i \in [k]$ . By assumption,  $T_i|X$  is a fully resolved triplet tree. Assume without loss of generality that  $T_i|X = x|yz$ . Then,  $T^{(q)}|X$  agrees with  $T_i|X$ , so  $T_i|X$  contributes +1 to  $f_q$ . On the other hand, both  $T^{(r)}|X$  and  $T^{(s)}|X$  disagree with  $T_i|X$ , so  $T_i|X$  contributes +1 to  $a_r$  and +1 to  $a_s$ . Furthermore, for any  $t \notin \{q, r, s\}$ ,  $T_i|X$  contributes nothing to  $f_t$  or  $a_t$ , since the triplet tree  $T^{(t)}|X$  is not fully resolved. Therefore, we have the following equalities.

$$\sum_{q=1}^d a_q = 2k \cdot |L| \quad (15)$$

$$\sum_{q=1}^d f_q = k \cdot |L| \quad (16)$$

Now suppose that for all  $q \in [d]$ ,  $f_q < a_q/2$ . This yields the following contradiction:

$$k \cdot |L| = \sum_{q=1}^d f_q < \frac{1}{2} \sum_{q=1}^d a_q = k \cdot |L|.$$

Here, the first equality follows from Equation (15) and the last equality follows from Equation (16). Thus, there must be some  $q \in [d]$  such that  $f_q \geq a_q/2$ .

Similarly, for the unrooted case, let  $L = \bigcup_{q,r \in [d], q \neq r} L_{qr}$ . Thus,  $L$  consists of those quartets that are unresolved in  $T$ , but resolved in  $T^{(qr)}$ , for some  $q, r \in [d]$ ,  $q \neq r$ . Equivalently,  $L$  consists of those quartets whose elements are leaves from four different neighboring subtrees of  $v$ .

Let  $X = \{w, x, y, z\}$  be a quartet in  $L$ . Assume that  $w \in \mathcal{L}(T(u_q, v))$ ,  $x \in \mathcal{L}(T(u_r, v))$ ,  $y \in \mathcal{L}(T(u_s, v))$ , and  $z \in \mathcal{L}(T(u_t, v))$ , where  $q, r, s, t$  must be distinct indices in  $[d]$ . Then,  $X$  is in  $L_q, L_r, L_s$ , and  $L_t$ .

Consider any  $i \in [k]$ . By assumption,  $T_i|X$  is a fully resolved quartet tree. Assume, without loss of generality, that  $T_i|X = wx|yz$ . Then,  $T^{(qr)}|X$  and  $T^{(st)}|X$  agree with  $T_i|X$ , so  $T_i|X$  contributes  $+1$  to  $f_{qr}$  and  $f_{st}$ , respectively. This double contribution is due to the symmetry of quartets. On the other hand,  $T^{(qs)}|X$ ,  $T^{(qt)}|X$ ,  $T^{(rs)}|X$ , and  $T^{(rt)}|X$  disagree with  $T_i|X$ , so  $T_i|X$  contributes  $+1$  to  $a_{qs}$ ,  $a_{qt}$ ,  $a_{rs}$ , and  $a_{rt}$ , respectively. Furthermore, if at least one of  $t_1, t_2 \notin \{q, r, s, t\}$ , then  $T_i|X$  contributes nothing to  $f_{t_1 t_2}$  or  $a_{t_1 t_2}$ , since the quartet tree  $T^{(t_1 t_2)}|X$  is not fully resolved. Therefore, similar to the rooted case, we have the following equalities.

$$\sum_{\substack{q, r \in [d] \\ q \neq r}} a_{qr} = 4k \cdot |L| \quad (17)$$

$$\sum_{\substack{q, r \in [d] \\ q \neq r}} f_{qr} = 2k \cdot |L| \quad (18)$$

Now suppose that for all  $q, r \in [d]$ ,  $q \neq r$ ,  $f_{qr} < a_{qr}/2$ . This yields the following contradiction:

$$2k \cdot |L| = \sum_{\substack{q, r \in [d] \\ q \neq r}} f_{qr} < \frac{1}{2} \sum_{\substack{q, r \in [d] \\ q \neq r}} a_{qr} = 2k \cdot |L|.$$

Here, the first equality follows from Equation (17) and the last equality follows from Equation (18). Thus, there must be some  $q, r \in [d]$ ,  $q \neq r$ , such that  $f_{q,r} \geq a_{qr}/2$ .  $\square$

**Lemma 5.2.** *Let  $\mathcal{P}$  be a profile for  $[k]$  over  $S$  consisting entirely of fully resolved rooted trees or fully resolved unrooted trees. Let  $T$  be a phylogeny for  $S$ ;  $T$  is rooted or unrooted according to whether  $\mathcal{P}$  consists of rooted or unrooted trees. Suppose  $T$  contains an unresolved node  $v$ , and suppose  $p \geq 2/3$ . Then, the following holds.*

- (i) *If  $T$  is rooted,  $v$  has a child  $u$  such that  $d^{(p)}(\widehat{T}, \mathcal{P}) \leq d^{(p)}(T, \mathcal{P})$ , where  $\widehat{T} = \text{PULL-OUT}(T, u)$ .*
- (ii) *If  $T$  is unrooted,  $v$  has two neighbors  $u_q$  and  $u_r$  such that  $d^{(p)}(\widehat{T}, \mathcal{P}) \leq d^{(p)}(T, \mathcal{P})$ , where  $\widehat{T} = \text{PULL-2-OUT}(T, u_q, u_r)$ .*

*Proof.* We will show that in the rooted case, for all  $q \in [d]$ ,

$$d^{(p)}(T^{(q)}, \mathcal{P}) = d^{(p)}(T, \mathcal{P}) - p \cdot f_q + (1 - p) \cdot a_q. \quad (19)$$

And, similarly, in the unrooted case, for all  $q, r \in [d]$ ,

$$d^{(p)}(T^{(qr)}, \mathcal{P}) = d^{(p)}(T, \mathcal{P}) - p \cdot f_{qr} + (1 - p) \cdot a_{qr}. \quad (20)$$

To verify this, consider any triplet or quartet  $X \in L_q$ . For every  $j$  such that  $T^{(q)}|X$  or  $T^{(qr)}|X$  is identical to  $T_j|X$ , the net change in the distance from  $\mathcal{P}$  is  $-p$ , since, for this  $X$ ,  $T_j$  contributes  $p$

to the distance to  $T$ , but contributes 0 to the distance to  $T^{(q)}$  or  $T^{(qr)}$ . For every  $j$  such that  $T^{(q)}|X$  or  $T^{(qr)}|X$  is different from  $T_j|X$ , the net change in the distance from  $\mathcal{P}$  is  $1 - p$ , since, for this  $X$ ,  $T_j$  contributes  $p$  to the distance to  $T$ , but contributes  $+1$  to the distance to  $T^{(q)}$  or  $T^{(qr)}$ .

Now, for the rooted case, choose an  $q^* \in [d]$  such that  $f_{q^*} \geq a_{q^*}/2$ ; for the unrooted case, choose two indices  $q^*, r^* \in [d]$ ,  $q^* \neq r^*$ , such that  $f_{q^*r^*} \geq a_{q^*r^*}/2$ . The existence of such a  $q^*$  (or  $q^*$  and  $r^*$ ) is guaranteed by Lemma 5.1. Then, Equation (19) and  $p \geq 2/3$  imply that  $d^{(p)}(T^{(q^*)}, \mathcal{P}) \leq d^{(p)}(T, \mathcal{P})$ . Similarly, Equation (20) and  $p \geq 2/3$  imply that  $d^{(p)}(T^{(q^*r^*)}, \mathcal{P}) \leq d^{(p)}(T, \mathcal{P})$ .  $\square$

*Proof of Theorem 5.2.* If  $\mathcal{P}$  consists of only fully-resolved trees, then any phylogeny  $T$  can be transformed into a fully-resolved tree  $T'$  such that  $d^{(p)}(T', \mathcal{P}) \leq d^{(p)}(T, \mathcal{P})$  by doing the following. First, let  $T' = T$ . Next, while  $T'$  contains an unresolved node, perform the following three steps:

1. Pick any unresolved node  $v$  in  $T'$ .
2. If  $T$  is rooted, find a child  $u$  of  $v$  such that  $d^{(p)}(\hat{T}, \mathcal{P}) \leq d^{(p)}(T, \mathcal{P})$ , where  $\hat{T} = \text{PULL-OUT}(T, u)$ .  
If  $T$  is unrooted, find two neighbors  $u_q, u_r$  of  $v$  such that  $d^{(p)}(\hat{T}, \mathcal{P}) \leq d^{(p)}(T, \mathcal{P})$ , where  $\hat{T} = \text{PULL-2-OUT}(T, u_q, u_r)$ .
3. Replace  $T'$  by  $\hat{T}$ .

Note that the existence of a node  $u$  such as the one required in Step 2 is guaranteed by Lemma 5.2. Thus, for  $p \geq 2/3$ , there always exists a fully-resolved median tree relative to  $d^{(p)}$ .  $\square$

The proof of Theorem 5.2 implies that if  $p > 2/3$  and the input trees are fully resolved, the median tree relative to  $d^{(p)}$  *must* be fully resolved. On the other hand, it is easy to show that when  $p \in [1/2, 2/3)$ , there are profiles of fully resolved trees whose median tree is only partially resolved.

## 6 Relationships among the metrics

We do not know whether the Hausdorff triplet or Hausdorff quartet distances are computable in polynomial time. Indeed, we suspect that, unlike their counterparts for partial rankings, this may not be possible. On the positive side, we show here that, in a broad range of cases, it is possible to obtain an approximation to the Hausdorff distance by exploiting its connection with parametric distance. As in the previous section, our results apply to both triplet and quartet distances. Our first result, which is proved later in this section, is as follows.

**Lemma 6.1.** *For every two phylogenies  $T_1$  and  $T_2$  over the same set of taxa,*

$$d_{\text{Haus}}(T_1, T_2) \geq |\mathcal{D}(T_1, T_2)| + \frac{2}{3} \cdot \max\{|\mathcal{R}_1(T_1, T_2)|, |\mathcal{R}_2(T_1, T_2)|\}.$$

An upper bound on  $d_{\text{Haus}}$  is obtained by assuming that  $T_1$  and  $T_2$  are refined so that the triplets (quartets) in  $\mathcal{R}_1(T_1, T_2)$ ,  $\mathcal{R}_2(T_1, T_2)$ , and  $\mathcal{U}(T_1, T_2)$  are resolved differently in each refinement. This gives us the following result, which we state without proof.

**Lemma 6.2.** *For every two phylogenies  $T_1$  and  $T_2$  over the same set of taxa,*

$$d_{\text{Haus}}(T_1, T_2) \leq |\mathcal{D}(T_1, T_2)| + |\mathcal{R}_1(T_1, T_2)| + |\mathcal{R}_2(T_1, T_2)| + |\mathcal{U}(T_1, T_2)|.$$

It is instructive to compare Lemmas 6.1 and 6.2 with the situation for partial rankings. The Hausdorff version of Kendall's tau is obtained by viewing each partial ranking as the set of all possible full rankings that can be obtained by refining it (that is, ordering elements within buckets). The distance is then the Hausdorff distance between the two sets, where the distance between two elements is the Kendall tau score. Critchlow [14] has given exact bounds on this distance measure, which allow it to be computed efficiently and to establish an equivalence with the parametric version of Kendall's tau defined in Section 5 [19]. To be precise, let  $L_1$  and  $L_2$  be two partial rankings. Re-using notation, let  $\mathcal{D}(L_1, L_2)$  be the set of all pairs that are ordered differently in  $L_1$  and  $L_2$ ,  $\mathcal{R}_1(L_1, L_2)$  be the set of pairs that are ordered in  $L_1$  but in the same bucket in  $L_2$ , and  $\mathcal{R}_2(L_1, L_2)$  be the set of pairs that are ordered in  $L_2$  but in the same bucket in  $L_1$ . Then, it can be shown that  $d_{\text{Haus}}(L_1, L_2) = |\mathcal{D}(L_1, L_2)| + \max\{|\mathcal{R}_1(L_1, L_2)|, |\mathcal{R}_2(L_1, L_2)|\}$  (see [14, 19]).

It seems unlikely that a similar simple expression can be obtained for Hausdorff triplet or quartet distance. There are at least two reasons for this. Let  $L_1$  and  $L_2$  be partial rankings. Then, it is possible to resolve  $L_1$  so that it disagrees with  $L_2$  in any pair in  $\mathcal{R}_2(L_1, L_2)$ . Similarly, there is a way to resolve  $L_2$  so that it disagrees with  $L_1$  in any pair in  $\mathcal{R}_1(L_1, L_2)$ . We have been unable to establish an analog of this property for trees; hence, the  $\frac{2}{3}$  factor in Lemma 6.1. The second reason is due to the properties of the set  $\mathcal{U}(L_1, L_2)$ . It can be shown that is one can refine rankings  $L_1$  and  $L_2$  in such a way that pairs of elements that are unresolved in both rankings are resolved the same way in the refinements. This seems impossible to do, in general, for trees and leads to the presence of  $|\mathcal{U}(T_1, T_2)|$  in Lemma 6.2.

The above observations prevent us from establishing equivalence between  $d_{\text{Haus}}$  and  $d^{(p)}$ , although they do not disprove equivalence either. In any event, the next result shows that when the number of triplets (quartets) that are unresolved in both trees is suitably small, equivalence *does* hold.

**Theorem 6.1.** *Let  $\beta$  be a positive real number. Then, for every  $p \in (0, 1]$ , Hausdorff distance and parametric distance are equivalent when restricted to pairs of trees  $(T_1, T_2)$  such that  $|\mathcal{U}(T_1, T_2)| \leq \beta(|\mathcal{D}(T_1, T_2)| + |\mathcal{R}_1(T_1, T_2)| + |\mathcal{R}_2(T_1, T_2)|)$ .*

*Proof.* By Proposition 5.1, it suffices to show that  $d_{\text{Haus}}$  is equivalent to  $d^{(2/3)}$ . Lemma 6.1 shows that  $d^{(2/3)}(T_1, T_2) \leq d_{\text{Haus}}(T_1, T_2)$ . Thus, we only need to show that, under our assumption about  $|\mathcal{U}(T_1, T_2)|$ , there is some  $c$  such that  $d_{\text{Haus}}(T_1, T_2) \leq c \cdot d^{(2/3)}(T_1, T_2)$ . The reader can verify that the result follows by choosing  $c = 3 + 3\beta$  and invoking Lemma 6.2.  $\square$

The remainder of this section is devoted to the proof of Lemma 6.1. The argument proceeds in two steps. First, we show that  $T_1$  can be refined so that it disagrees with  $T_2$  in at least two thirds of the triplets (quartets) in  $\mathcal{R}_2(T_1, T_2)$ . Next, we show the existence of an analogous refinement of  $T_2$ . Note that the triplets (quartets) in  $\mathcal{D}(T_1, T_2)$  are resolved differently in any refinements of  $T_1$  and  $T_2$ . This gives lower bounds for both arguments in the outer max of the definition of  $d_{\text{Haus}}(T_1, T_2)$  (Equation 2) and yields the lemma.



Let  $v$  be a node in  $T_1$ . If  $T_1$  is rooted, then, as in Section 5, let  $u_1, \dots, u_d$  denote the children of  $v$  in  $T_1$  and  $T_1^{(q)}$  denote  $\text{PULL-OUT}(T, u_q)$ . Define  $\mathcal{M}_q(v)$  to be the set of all triplets  $X \in \mathcal{R}_2(T_1, T_2)$  such that (i) the lca of  $X$  in  $T_1$  is  $v$  and (ii)  $T_1|X$  is unresolved but  $T_1^{(q)}|X$  is fully resolved. Let  $\mathcal{M}(v) = \bigcup_{q=1}^d \mathcal{M}_q(v)$ . Thus,  $\mathcal{M}(v)$  is the set of triplets associated with  $v$  that are resolved in  $T_2$  but not in  $T_1$ .

If  $T_1$  is unrooted,  $u_1, \dots, u_d$  denote the neighbors of  $v$  in  $T_1$  and  $T_1^{(qr)}$  denotes  $\text{PULL-2-OUT}(T_1, u_{qr})$ , where  $\text{PULL-2-OUT}$  is the function defined in Section 5. Define  $\mathcal{M}_{qr}(v)$  to be the set of all quartets  $X \in \mathcal{R}_2(T_1, T_2)$  such that (i)  $T_1|X$  is a fan, (ii) the paths between any two distinct pairs of taxa in  $X$  meet at  $v$ , and (iii)  $T_1|X$  is unresolved but  $T_1^{(qr)}|X$  is fully resolved. Let  $\mathcal{M}(v) = \bigcup_{q,r \in [d], q \neq r} \mathcal{M}_{qr}(v)$ . Thus,  $\mathcal{M}(v)$  is the set of quartets associated with  $v$  that are resolved in  $T_2$  but not in  $T_1$ .

Define the following two sets for the rooted case.

$$F_q = \{X \in \mathcal{M}_q(v) : T_2|X \text{ agrees with } T_1^{(q)}|X\} \quad (21)$$

$$A_q = \{X \in \mathcal{M}_q(v) : T_2|X \text{ disagrees with } T_1^{(q)}|X\}. \quad (22)$$

Define the following two sets for the unrooted case.

$$F_{qr} = \{X \in \mathcal{M}_{qr}(v) : T_2|X \text{ agrees with } T_1^{(qr)}|X\} \quad (23)$$

$$A_{qr} = \{X \in \mathcal{M}_{qr}(v) : T_2|X \text{ disagrees with } T_1^{(qr)}|X\}. \quad (24)$$

The next result is, in a sense, a counterpart to Lemma 5.1.

**Lemma 6.3.** *For the rooted case, there exists an index  $q \in [d]$  such that  $|A_q| \geq 2|F_q|$ . For the unrooted case, there exist two indices  $q, r \in [d]$ ,  $q \neq r$ , such that  $|A_{qr}| \geq 2|F_{qr}|$ .*

*Proof.* We start with the rooted case. Consider any triplet  $X = \{x, y, z\}$  in  $\mathcal{M}(v)$ . Assume that  $x \in \mathcal{L}(T_1(u_q))$ ,  $y \in \mathcal{L}(T_1(u_r))$ , and  $z \in \mathcal{L}(T_1(u_s))$ , where  $q, r, s$  must be distinct indices in  $[d]$ . Thus,  $X$  is in  $\mathcal{M}_q(v)$ ,  $\mathcal{M}_r(v)$ , and  $\mathcal{M}_s(v)$ .

By definition of  $\mathcal{M}(v)$ ,  $T_2|X$  is a fully resolved triplet tree. Assume that  $T_2|X = x|yz$ . Then,  $T_1^{(q)}|X$  agrees with  $T_2|X$ , so  $X$  contributes exactly one element to  $F_q$ . On the other hand, both  $T_1^{(r)}|X$  and  $T_1^{(s)}|X$  disagree with  $T_2|X$ , so  $X$  contributes exactly one element to  $A_r$  and one element to  $A_s$ . Furthermore, for any  $t \notin \{q, r, s\}$ ,  $X$  contributes nothing to  $F_t$  or  $A_t$ , since the triplet tree  $T_1^{(t)}|X$  is not fully resolved. Therefore, we have that

$$\sum_{q=1}^d |A_q| = 2 \cdot |\mathcal{M}(v)| \quad \text{and} \quad \sum_{q=1}^d |F_q| = |\mathcal{M}(v)|. \quad (25)$$

Assume that for all  $q \in [d]$ ,  $|F_q| > |A_q|/2$ . This and (25) imply that

$$|\mathcal{M}(v)| = \sum_{q=1}^d |F_q| > \frac{1}{2} \sum_{q=1}^d |A_q| = |\mathcal{M}(v)|,$$

a contradiction.

We now consider the unrooted case. Consider any quartet  $X = \{w, x, y, z\}$  in  $\mathcal{M}(v)$ . Assume that  $w \in \mathcal{L}(T_1(u_q, v))$ ,  $x \in \mathcal{L}(T_1(u_r, v))$ ,  $y \in \mathcal{L}(T_1(u_s, v))$ , and  $z \in \mathcal{L}(T_1(u_t, v))$ , where  $q, r, s, t$  must be distinct indices in  $[d]$ . Thus,  $X$  is in  $\mathcal{M}_{qr}(v)$ ,  $\mathcal{M}_{qs}(v)$ ,  $\mathcal{M}_{qt}(v)$ ,  $\mathcal{M}_{rs}(v)$ ,  $\mathcal{M}_{rt}(v)$  and  $\mathcal{M}_{st}(v)$ .

By definition of  $\mathcal{M}(v)$ ,  $T_2|X$  is a fully resolved quartet tree. Assume that  $T_2|X = wx|yz$ . Then,  $T_1^{(qr)}|X$  and  $T_1^{(st)}|X$  agree with  $T_2|X$ , so  $X$  contributes exactly one element to  $F_{qr}$  and  $F_{st}$ . On the other hand,  $T_1^{(qs)}|X$ ,  $T_1^{(qt)}|X$ ,  $T_1^{(rs)}|X$  and  $T_1^{(rt)}|X$  disagree with  $T_2|X$ , so  $X$  contributes exactly one element to  $A_{qs}$ ,  $A_{qt}$ ,  $A_{rs}$  and  $A_{rt}$ , respectively. Furthermore, for any  $j_1$  and  $j_2 \notin \{q, r, s, t\}$ ,  $X$  contributes nothing to  $F_{j_1 j_2}$  or  $A_{j_1 j_2}$ , since the quartet tree  $T_1^{(j_1 j_2)}|X$  is not fully resolved. Therefore, we have that

$$\sum_{\substack{q, r \in [d] \\ q \neq r}} |A_{qr}| = 4 \cdot |\mathcal{M}(v)| \quad \text{and} \quad \sum_{\substack{q, r \in [d] \\ q \neq r}} |F_{qr}| = 2 \cdot |\mathcal{M}(v)|. \quad (26)$$

Assume that for all  $q, r \in [d]$ ,  $|F_{qr}| > |A_{qr}|/2$ . This and (26) imply that

$$2 \cdot |\mathcal{M}(v)| = \sum_{\substack{q, r \in [d] \\ q \neq r}} |F_{qr}| > \frac{1}{2} \sum_{\substack{q, r \in [d] \\ q \neq r}} |A_{qr}| = 2 \cdot |\mathcal{M}(v)|,$$

a contradiction. □

*Proof of Lemma 6.1.* Define the following functions. For any two phylogenies  $T_1, T_2$  over  $S$ , let

$$d_{H1}(T_1, T_2) = \max_{t_1 \in \mathcal{F}(T_1)} \min_{t_2 \in \mathcal{F}(T_2)} d(t_1, t_2), \quad (27)$$

$$d_{H2}(T_1, T_2) = \max_{t_2 \in \mathcal{F}(T_2)} \min_{t_1 \in \mathcal{F}(T_1)} d(t_1, t_2). \quad (28)$$

We show that

$$d_{H1}(T_1, T_2) \geq |\mathcal{D}(T_1, T_2)| + \frac{2}{3} \cdot |\mathcal{R}_2(T_1, T_2)| \quad (29)$$

$$d_{H2}(T_1, T_2) \geq |\mathcal{D}(T_1, T_2)| + \frac{2}{3} \cdot |\mathcal{R}_1(T_1, T_2)|. \quad (30)$$

Since  $d_{\text{Haus}}(T_1, T_2) = \max\{d_{H1}(T_1, T_2), d_{H2}(T_1, T_2)\}$ , this proves Lemma 6.1.

By symmetry, it suffices to prove Inequality (29). Our argument relies on two observations. First, note that if  $T'_1$  is a refinement of  $T_1$  (but possibly not a full refinement), then,  $d_{H1}(T_1, T_2) \geq d_{H1}(T'_1, T_2)$ . This holds because  $\mathcal{F}(T'_1) \subseteq \mathcal{F}(T_1)$ . Second, for any two phylogenies  $T_1$  and  $T_2$ ,  $d_{H1}(T_1, T_2) \geq |\mathcal{D}(T_1, T_2)|$ . This holds because for any  $t_1 \in \mathcal{F}(T_1)$ ,  $t_2 \in \mathcal{F}(T_2)$ , we have that  $\mathcal{D}(T_1, T_2) \subseteq \mathcal{D}(t_1, t_2)$ , and (by definition)  $d(t_1, t_2) = |\mathcal{D}(t_1, t_2)|$ .

By the preceding observations, if we prove that it is possible to construct a refinement  $T'_1$  of  $T_1$  such that  $|\mathcal{D}(T'_1, T_2)| \geq |\mathcal{D}(T_1, T_2)| + \frac{2}{3}|\mathcal{R}_2(T_1, T_2)|$ , then Inequality (29) follows. The idea is to find a refinement  $T'_1$  of  $T_1$  such that for at least two-thirds of the triplets or quartets  $X \in \mathcal{R}_2(T_1, T_2)$ , we have that  $T'_1|X \neq T_2|X$ . To obtain the desired refinement of  $T_1$ , we initially set  $T'_1 = T_1$  and then perform the following steps while they apply:

1. Pick an unresolved node  $v$  in  $T'_1$  such that  $\mathcal{M}'(v) \neq \emptyset$ , where  $\mathcal{M}'(v)$  is the set of triplets (quartets) associated with  $v$  that are resolved in  $T_2$  but not in  $T'_1$ . In the rooted case, let  $u_1, \dots, u_d$  be the children of  $v$ ; in the unrooted case, let  $u_1, \dots, u_d$  be the neighbors of  $v$ .
2. For rooted trees, find a  $q \in [d]$  such that  $|A_q| \geq 2|F_q|$  (such a  $q$  exists by Lemma 6.3). For unrooted trees, find  $q, r \in [d]$  such that  $|A_{qr}| \geq 2|F_{qr}|$  (such  $q, r$  exist by Lemma 6.3).
3. In the rooted case, set  $T'_1 = \text{PULL-OUT}(T'_1, u_q)$ ; in the unrooted case, set  $T'_1 = \text{PULL-2-OUT}(T'_1, u_q, u_r)$ .

When this algorithm terminates,  $\mathcal{M}'(v) = \emptyset$  for every  $v \in \mathcal{V}(T'_1)$ . Thus,  $\mathcal{R}_2(T'_1, T_2) = \emptyset$ . Furthermore, the choice of  $q$  (or  $q_1$  and  $q_2$ ) in step (2) guarantees that  $|\mathcal{D}(T'_1, T_2)| \geq |\mathcal{D}(T_1, T_2)| + \frac{2}{3} \cdot |\mathcal{R}_2(T_1, T_2)|$ .  $\square$

## 7 Computing parametric triplet distance

In this section we show that the parametric triplet distance (PTD),  $d^{(p)}$ , between two phylogenetic trees  $T_1$  and  $T_2$  over the same set of  $n$  taxa can be computed in  $O(n^2)$  time.

Before we outline our PTD algorithm, we need some notation. Let  $T$  be a rooted phylogenetic tree. Then,  $R(T)$  denotes the set of all triplets that are resolved in  $T$  and  $U(T)$  denotes the set of all triplets that are unresolved in  $T$ .

The next proposition is easily proved.

**Proposition 7.1.** *For any two phylogenies  $T_1, T_2$  over the same set of taxa,*

- (i)  $|\mathcal{R}_1(T_1, T_2)| + |\mathcal{U}(T_1, T_2)| = |U(T_2)|$
- (ii)  $|\mathcal{R}_2(T_1, T_2)| + |\mathcal{U}(T_1, T_2)| = |U(T_1)|$ ,
- (iii)  $|\mathcal{S}(T_1, T_2)| + |\mathcal{D}(T_1, T_2)| + |\mathcal{R}_1(T_1, T_2)| = |R(T_1)|$ .

By Prop. 7.1 and Eqn. (1), the parametric distance between  $T_1$  and  $T_2$  can be expressed as

$$d^{(p)}(T_1, T_2) = |R(T_1)| - |\mathcal{S}(T_1, T_2)| + p \cdot (|U(T_1)| - |U(T_2)|) + (2p - 1) \cdot |\mathcal{R}_1(T_1, T_2)|. \quad (31)$$

Our PTD algorithm proceeds as follows. After an initial  $O(n^2)$  preprocessing step (Section 7.1), the algorithm computes  $|R(T_1)|$ ,  $|U(T_1)|$  and  $|U(T_2)|$  using a  $O(n)$ -time procedure (Section 7.2). Next, it computes  $|\mathcal{S}(T_1, T_2)|$  and  $|\mathcal{R}_1(T_1, T_2)|$ . As described in Sections 7.3 and 7.4, this takes  $O(n^2)$  time. Then, it uses these values to compute  $d^{(p)}(T_1, T_2)$ , in  $O(1)$  time, via Equation (31). To summarize, we have the following result.

**Theorem 7.1.** *The parametric triplet distance  $d^{(p)}(T_1, T_2)$  for two rooted phylogenetic trees  $T_1$  and  $T_2$  over the same set of  $n$  taxa can be computed in  $O(n^2)$  time.*

In the rest of this section we use the following notation. We write  $rt(T)$  to denote the root node of a tree  $T$ . Let  $v$  be a node in  $T$ . Then,  $pa(v)$  denotes the parent of  $v$  in  $T$  and  $Ch(v)$  is the set of children of  $v$ . We write  $\overline{T(v)}$  to denote the tree obtained by deleting  $T(v)$  from  $T$ , as well as the edge from  $v$  to its parent, if such an edge exists.

## 7.1 The preprocessing step

The purpose of the preprocessing step is to calculate and store the following four quantities for every pair  $(u, v)$ , where  $u \in \mathcal{V}(T_1)$  and  $v \in \mathcal{V}(T_2)$ :  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$ ,  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(\overline{T_2(v)})|$ ,  $|\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(T_2(v))|$ , and  $|\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|$ . These values are stored in a table so that any value can be accessed in  $O(1)$  time by subsequent steps of the PTD algorithm.

**Lemma 7.1.** *The values  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$ ,  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(\overline{T_2(v)})|$ ,  $|\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(T_2(v))|$ , and  $|\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|$  can be collectively computed for every pair of nodes  $(u, v)$ , where  $u \in \mathcal{V}(T_1)$  and  $v \in \mathcal{V}(T_2)$ , in  $O(n^2)$  time.*

*Proof.* We first observe that for each  $u \in \mathcal{V}(T_1)$ , the value  $|\mathcal{L}(T_1(u))|$  can be computed in  $O(n)$  time by a simple post order traversal of  $T_1$ . The same holds for tree  $T_2$ .

Consider the value  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$ . We consider three cases.

1. If  $u$  and  $v$  are both leaf nodes then computing  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$  is trivial.
2. If  $u$  is a leaf node, but  $v$  is not a leaf node, then

$$|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))| = \sum_{x \in \text{Ch}(v)} |\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(x))|.$$

3. If  $u$  is not a leaf node, then

$$|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))| = \sum_{x \in \text{Ch}(u)} |\mathcal{L}(T_1(x)) \cap \mathcal{L}(T_2(v))|.$$

We compute the value  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$ , for every pair  $(u, v)$ , using an interleaved post order traversal of  $T_1$  and  $T_2$ . This traversal works as follows: For each node  $u$  in a post order traversal of  $T_1$ , we consider each node  $v$  in a post order traversal of  $T_2$ . This ensures that when the intersection sizes for a pair of nodes is computed, the set intersection sizes for all pairs of their children have already been computed. The total time complexity for computing the required values in this way can be bounded as follows. For a pair of nodes  $u$  and  $v$  from  $T_1$  and  $T_2$  respectively, the value  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$  can be computed in  $O(|\text{Ch}(u)| + |\text{Ch}(v)|)$  time and all the remaining three set intersection values in  $O(1)$  time. Summing this over all possible pairs of edges, we get a total time of  $O(\sum_{u \in \mathcal{V}(T_1)} \sum_{v \in \mathcal{V}(T_2)} (|\text{Ch}(u)| + |\text{Ch}(v)|))$ , which is  $O(n^2)$ .

Once the value  $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$  has been computed for every pair  $(u, v)$ , the remaining quantities we seek can be computed using the following relations.

$$\begin{aligned} |\mathcal{L}(T_1(u)) \cap \mathcal{L}(\overline{T_2(v)})| &= |\mathcal{L}(T_1(u))| - |\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|, \\ |\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(T_2(v))| &= |\mathcal{L}(T_2(v))| - |\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|, \quad \text{and} \\ |\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})| &= n - (|\mathcal{L}(T_1(u))| + |\mathcal{L}(T_2(v))| - |\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|). \end{aligned}$$

Thus, each of these values can be computed in  $O(1)$  time, for a total of  $O(n^2)$ .  $\square$

We store these  $O(n^2)$  values in an array indexed by  $u$  and  $v$ , for each  $u \in \mathcal{V}(T_1)$  and  $v \in \mathcal{V}(T_2)$ . This enables constant time insertion and look-up of any stored value, when the two relevant nodes are given.

## 7.2 Computing $|R(T_1)|$ , $|U(T_1)|$ , and $|U(T_2)|$

Here we prove the following result.

**Lemma 7.2.** *Given a rooted phylogenetic tree  $T$  over  $n$  leaves, the values  $|R(T)|$  and  $|U(T)|$  can be computed in  $O(n)$  time.*

Thus,  $|R(T_1)|$ ,  $|U(T_1)|$  and  $|U(T_2)|$  can all be computed in  $O(n)$  time.

To prove Lemma 7.2, we need some terminology and an auxiliary result. Let  $e = (v, pa(v))$  be any internal edge in  $T$ . Consider any two leaves  $x, y$  from  $\mathcal{L}(T(v))$ , and any leaf  $z$  from  $\mathcal{L}(\overline{T(v)})$ . Then, the triplet  $\{x, y, z\}$  must appear resolved as  $xy|z$  in  $T$ ; we say that the triplet tree  $xy|z$  is *induced* by the edge  $(v, pa(v))$ . Note that the same resolved triplet tree may be induced by multiple edges in  $T$ . We say that the triplet tree  $xy|z$  is *strictly induced* by the edge  $\{v, pa(v)\}$  if  $xy|z$  is induced by  $(v, pa(v))$  and, additionally,  $x \in \mathcal{L}(T(v_1))$  and  $y \in \mathcal{L}(T(v_2))$  for some  $v_1, v_2 \in Ch(v)$  such that  $v_1 \neq v_2$ .

**Lemma 7.3.** *Given a tree  $T$  and a triplet  $X$ , if  $T|X$  is fully resolved then  $T|X$  is strictly induced by exactly one edge in  $T$ .*

*Proof.* Let  $X = \{a, b, c\}$ . Without loss of generality, assume that  $T|X = ab|c$ . If  $v$  denotes the lca of  $a$  and  $b$  in  $T$ , the edge  $\{v, pa(v)\}$  must induce  $ab|c$ . Moreover,  $v$  must be the only node in  $T$  for which there exist nodes  $v_1, v_2 \in Ch(v)$  such that  $a \in \mathcal{L}(T(v_1))$  and  $b \in \mathcal{L}(T(v_2))$ . Thus, there is exactly one edge in  $T$  that strictly induces  $T|X$ .  $\square$

*Proof of Lemma 7.2.* Since  $|R(T)| + |U(T)| = \binom{n}{3}$ , given  $|R(T)|$ , the value  $|U(T)|$  can be computed in  $O(1)$  additional time. Thus, we only need to show that the value of  $|R(T)|$  can be computed in  $O(n)$  time.

The first step is to traverse the tree  $T$  in post order to compute the values  $\alpha_v = |\mathcal{L}(T(v))|$  and  $\beta_v = n - \alpha_v$  at each node  $v \in \mathcal{V}(T)$ . This takes  $O(n)$  time.

For any  $v \in \mathcal{V}(T) \setminus \{rt(T)\}$ , let  $\phi(v)$  denote the number of triplets that are strictly induced by the edge  $\{v, pa(v)\}$  in tree  $T$ . Observe that any triplet that is strictly induced by an edge in  $T$  must be fully resolved in  $T$ . Thus, Lemma 7.3 implies that the sum of  $\phi(v)$  over all internal nodes  $v \in \mathcal{V}(T) \setminus \{rt(T)\}$  yields the value  $|R(T)|$ . We now show how to compute the value of  $\phi(v)$ .

Let  $X = \{a, b, c\}$  be a triplet that is counted in  $\phi(v)$ . And, without loss of generality, let  $T_1|X = ab|c$ . It can be verified that  $X$  must satisfy the following two conditions: (i)  $a, b \in \mathcal{L}(T(v))$  and  $c \in \mathcal{L}(\overline{T(v)})$ , and (ii) there does not exist any  $x \in Ch(v)$  such that  $a, b \in \mathcal{L}(T(x))$ . The number of triplets that satisfy condition (i) is  $\binom{\alpha_v}{2} \cdot \beta_v$ , and the number of triplets that satisfy condition (i), but not condition (ii) is exactly  $\sum_{x \in Ch(v)} \binom{\alpha_x}{2} \cdot \beta_v$ . Thus,  $\phi(v) = \gamma_v - \sum_{x \in Ch(v)} \binom{\alpha_x}{2} \cdot \beta_v$ .

Computing  $\phi(v)$  requires  $O(|Ch(v)|)$  time; hence, the time complexity for computing  $|R(T)|$  is  $O(\sum_{v \in \mathcal{V}(T)} |Ch(v)|)$ , which is  $O(n)$ .  $\square$

## 7.3 Computing $|\mathcal{S}(T_1, T_2)|$

We now describe an  $O(n^2)$  time algorithm to compute the size of the set  $\mathcal{S}(T_1, T_2)$  of shared triplets; that is, triplets that are fully and identically resolved in  $T_1$  and  $T_2$ .

For any  $u \in \mathcal{V}(T_1) \setminus (rt(T_1) \cup \mathcal{L}(T_1))$  and  $v \in \mathcal{V}(T_2) \setminus (rt(T_2) \cup \mathcal{L}(T_2))$ , let  $s(u, v)$  denote the number of identical triplet trees strictly induced by edge  $\{u, pa(u)\}$  in  $T_1$  and edge  $\{v, pa(v)\}$  in  $T_2$ . We have the following result.

**Lemma 7.4.** *Given  $T_1$  and  $T_2$ , we have,*

$$|\mathcal{S}(T_1, T_2)| = \sum_{\substack{u \in \mathcal{V}(T_1) \setminus (rt(T_1) \cup \mathcal{L}(T_1)), \\ v \in \mathcal{V}(T_2) \setminus (rt(T_2) \cup \mathcal{L}(T_2))}} s(u, v). \quad (32)$$

*Proof.* Consider any triplet  $X \in \mathcal{S}(T_1, T_2)$ . Since  $T_1|X$  is fully resolved and  $T_1|X = T_2|X$  then, by Lemma 7.3, there exists exactly one node  $u \in \mathcal{V}(T_1) \setminus rt(T_1)$  and one node  $v \in \mathcal{V}(T_2) \setminus rt(T_2)$  such that the edge  $\{u, pa(u)\}$  strictly induces  $T_1|X$  in  $T_1$ , and edge  $\{v, pa(v)\}$  strictly induces  $T_2|X$  in  $T_2$ . Additionally, neither  $u$  nor  $v$  can be leaf nodes in  $T_1$  and  $T_2$  respectively. Thus,  $X$  would be counted exactly once in the right-hand side of Equation (32) in the value  $s(u, v)$ . Moreover, by the definition of  $s(u, v)$ , any triplet tree that is counted on the right-hand side of Equation (32) algorithm must belong to the set  $\mathcal{S}(T_1, T_2)$ . The Lemma follows.  $\square$

The following lemma shows how to compute the value of  $s(u, v)$  using the values computed in the preprocessing step.

**Lemma 7.5.** *Given any  $u \in \mathcal{V}(T_1) \setminus (rt(T_1) \cup \mathcal{L}(T_1))$  and  $v \in \mathcal{V}(T_2) \setminus (rt(T_2) \cup \mathcal{L}(T_2))$ ,  $s(u, v)$  can be computed in  $O(|Ch(u)| \cdot |Ch(v)|)$  time.*

*Proof.* We will show that  $s(u, v) = n_1(u, v) - n_2(u, v) - n_3(u, v) + n_4(u, v)$ , where

$$\begin{aligned} n_1(u, v) &= \binom{|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|}{2} \cdot |\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|, \\ n_2(u, v) &= \sum_{x \in Ch(u)} \binom{|\mathcal{L}(T_1(x)) \cap \mathcal{L}(T_2(v))|}{2} \cdot |\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|, \\ n_3(u, v) &= \sum_{x \in Ch(v)} \binom{|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(x))|}{2} \cdot |\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|, \quad \text{and} \\ n_4(u, v) &= \sum_{x \in Ch(u)} \sum_{y \in Ch(v)} \binom{|\mathcal{L}(T_1(x)) \cap \mathcal{L}(T_2(y))|}{2} \cdot |\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|. \end{aligned}$$

Consider any triplet tree,  $ab|c$ , counted in  $s(u, v)$ . It can be verified that  $ab|c$  must satisfy the following three conditions: (i)  $a, b \in \mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))$  and  $c \in \mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})$ , (ii) there does not exist any  $x \in Ch(u)$  such that  $a, b \in \mathcal{L}(T_1(x))$ , and (iii) there does not exist any  $x \in Ch(v)$  such that  $a, b \in \mathcal{L}(T_2(x))$ . Moreover, observe that any triplet tree  $ab|c$  that satisfies these three conditions is counted in  $s(u, v)$ . Therefore,  $s(u, v)$  is exactly the number of triplet trees that satisfy all three conditions (i), (ii) and (iii).

The number of triplet trees that satisfy condition (i) is given by  $n_1(u, v)$ . Some of the triplet trees that satisfy condition (i) may not satisfy conditions (ii) or (iii); these must not be counted in

**procedure**  $\mathcal{S}(T_1, T_2)$   
1: **for** each internal node  $u \in \mathcal{V}(T_1) \setminus \text{rt}(T_1)$  **do**  
2:   **for** each internal node  $v \in \mathcal{V}(T_2) \setminus \text{rt}(T_2)$  **do**  
3:     Compute  $s(u, v)$ .  
4: **return** the sum of all computed  $s(\cdot, \cdot)$ .

Figure 1: Computing  $|\mathcal{S}(T_1, T_2)|$

$s(u, v)$ . The value  $n_2(u, v)$  is exactly the number of triplet trees that satisfy condition (i) but not condition (ii). Similarly,  $n_3(u, v)$  is exactly the number of triplet trees that satisfy condition (i) but not (iii). Thus, the second and third terms must be subtracted from the first term. However, there may be triplet trees that satisfy condition (i) but neither (ii) nor (iii), and, consequently, get subtracted in both the second and third terms. In order to adjust for these, the value  $n_4(u, v)$  counts exactly those triplet trees that satisfy condition (i) but not (ii) and (iii).  $\square$

A summary of our algorithm to compute  $|\mathcal{S}(T_1, T_2)|$  appears in Figure 1.

**Lemma 7.6.** *Given two rooted phylogenetic trees  $T_1$  and  $T_2$  on the same  $n$  leaves, the value  $|\mathcal{S}(T_1, T_2)|$  can be computed in  $O(n^2)$  time.*

*Proof.* By Lemma 7.4, the algorithm of Figure 1 computes the value  $|\mathcal{S}(T_1, T_2)|$  correctly. We now analyze its complexity. The running time of the algorithm is dominated by the complexity of computing the value  $s(u, v)$  for each pair of internal nodes  $u \in \mathcal{V}(T_1)$  and  $v \in \mathcal{V}(T_2)$ . According to Lemma 7.5, the value  $s(u, v)$  can be computed in  $O(|\text{Ch}(u)| \cdot |\text{Ch}(v)|)$  time. Thus, the total time complexity of the algorithm is  $O(\sum_{u \in \mathcal{V}(T_1)} \sum_{v \in \mathcal{V}(T_2)} |\text{Ch}(u)| \cdot |\text{Ch}(v)|)$ , which is  $O(n^2)$ .  $\square$

## 7.4 Computing $|\mathcal{R}_1(T_1, T_2)|$

Next, we describe an  $O(n^2)$ -time algorithm that computes the cardinality of the set  $\mathcal{R}_1(T_1, T_2)$  of triplets that are resolved only in tree  $T_1$ . First, we need a definition. Let  $X$  be a triplet that is unresolved in  $T_2$ . Let  $v$  be the least common ancestor (lca) of  $X$  in  $T_2$ . We say that  $X$  is *associated* with  $v$ . Observe that node  $v$  must be internal and unresolved. Note also that  $X$  is associated with exactly one node in  $T_2$ .

For any  $u \in \mathcal{V}(T_1) \setminus (\text{rt}(T_1) \cup \mathcal{L}(T_1))$  and  $v \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)$ , let  $r_1(u, v)$  denote the number of triplets  $X$  such that  $T_1|X$  is strictly induced by edge  $\{u, \text{pa}(u)\}$  in  $T_1$ , and  $X$  is associated with the node  $v$  in  $T_2$ .

The triplets counted in  $r_1(u, v)$  must be resolved in  $T_1$  but unresolved in  $T_2$ . Our algorithm computes the value  $|\mathcal{R}_1(T_1, T_2)|$  by computing, for each  $u \in \mathcal{V}(T_1) \setminus (\text{rt}(T_1) \cup \mathcal{L}(T_1))$  and  $v \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)$ , the value  $r_1(u, v)$ . We claim that the sum of all the computed  $r_1(u, v)$ 's yields the value  $|\mathcal{R}_1(T_1, T_2)|$ .

**Lemma 7.7.** *Given  $T_1$  and  $T_2$ , we have,*

$$|\mathcal{R}(T_1, T_2)| = \sum_{\substack{u \in \mathcal{V}(T_1) \setminus (\text{rt}(T_1) \cup \mathcal{L}(T_1)), \\ v \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)}} r_1(u, v). \quad (33)$$

*Proof.* Consider any triplet  $X \in \mathcal{R}_1(T_1, T_2)$ . By Lemma 7.3, there exists exactly one node  $u \in \mathcal{V}(T_1) \setminus rt(T_1)$  such that the edge  $\{u, pa(u)\}$  strictly induces  $T_1|X$  in  $T_1$ . Also observe that there must be exactly one unresolved node  $v \in \mathcal{V}(T_2)$  with which  $X$  is associated. Additionally, neither  $u$  nor  $v$  can be leaf nodes in  $T_1$  and  $T_2$  respectively. Thus,  $X$  would be counted exactly once in the right-hand side of Equation (33); in the value  $r_1(u, v)$ . Moreover, by the definition of  $r_1(u, v)$ , any triplet that is counted in the right-hand side of Equation (33) must belong to the set  $\mathcal{R}_1(T_1, T_2)$ . The lemma follows.  $\square$

Given a path  $u_1, u_2, \dots, u_k$ , where  $k \geq 2$ , in tree  $T_1$  such that  $u_k$  is an internal node and  $u_1$  is an ancestor of  $u_k$ , let  $\gamma(u_1, u_k, v)$  denote the number of triplets  $X$  such that  $T_1|X$  is induced by every edge  $\{u_{i-1}, u_i\}$ , for  $2 \leq i \leq k$ , in  $T_1$  and  $X$  is associated with node  $v$  in  $T_2$ .

The following lemma shows how the value of  $r_1(u, v)$  can be computed by first computing certain  $\gamma(\cdot, \cdot, \cdot)$  values.

**Lemma 7.8.** *For any  $u \in \mathcal{V}(T_1) \setminus (rt(T_1) \cup \mathcal{L}(T_1))$  and  $v \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)$ ,*

$$r_1(u, v) = \gamma(pa(u), u, v) - \sum_{x \in Ch(u)} \gamma(pa(u), x, v).$$

*Proof.* Let  $X = \{a, b, c\}$  be a triplet that is counted in  $r_1(u, v)$ . And, without loss of generality, let  $T_1|X = ab|c$ . It can be verified that  $X$  must satisfy the following three conditions: (i)  $X$  must be associated with  $v$  in  $T_2$ , (ii)  $a, b \in \mathcal{L}(T_1(u))$  and  $c \in \mathcal{L}(\overline{T_1(u)})$ , and (iii) there must not exist any  $x \in Ch(u)$  such that  $a, b \in \mathcal{L}(T_1(x))$ . Moreover, observe that if there exists a triplet  $X = \{a, b, c\}$  that satisfies these three conditions, then  $X$  will be counted in  $r_1(u, v)$ ; these three conditions are thus necessary and sufficient.

Now observe that  $\gamma(pa(u), u, v)$  counts exactly those triplets that satisfy conditions (i) and (ii), while  $\sum_{x \in Ch(u)} \gamma(pa(u), x, v)$  counts exactly those triplets that satisfy conditions (i) and (ii), but not condition (iii). The lemma follows immediately.  $\square$

To compute the value of  $\gamma(\cdot, \cdot, \cdot)$  efficiently we use the following lemma.

**Lemma 7.9.** *Consider a path  $u_1, u_2, \dots, u_k$ , where  $k \geq 2$ , in tree  $T_1$  such that  $u_k$  is an internal node and  $u_1$  is an ancestor of  $u_k$ . And let  $v \in \mathcal{V}(T_2)$  be an internal unresolved node. Then,*

$$\gamma(u_1, u_k, v) = n_1(u_1, u_k, v) - n_2(u_1, u_k, v) - n_3(u_1, u_k, v) - n_4(u_1, u_k, v),$$

where

$$\begin{aligned} n_1(u_1, u_k, v) &= \binom{|\mathcal{L}(T_2(v)) \cap \mathcal{L}(T_1(u_k))|}{2} \cdot |\mathcal{L}(T_2(v)) \cap \mathcal{L}(\overline{T_1(u_2)})|, \\ n_2(u_1, u_k, v) &= \sum_{x \in Ch(v)} \binom{|\mathcal{L}(T_2(x)) \cap \mathcal{L}(T_1(u_k))|}{2} \cdot |\mathcal{L}(T_2(x)) \cap \mathcal{L}(\overline{T_1(u_2)})|, \\ n_3(u_1, u_k, v) &= \sum_{x \in Ch(v)} \binom{|\mathcal{L}(T_1(u_k)) \cap \mathcal{L}(T_2(x))|}{2} \cdot (|\mathcal{L}(T_2(v)) \cap \mathcal{L}(\overline{T_1(u_2)})| - |\mathcal{L}(T_2(x)) \cap \mathcal{L}(\overline{T_1(u_2)})|), \end{aligned}$$



and

$$n_4(u_1, u_k, v) = \sum_{x \in \text{Ch}(v)} |\mathcal{L}(T_2(x)) \cap \mathcal{L}(T_1(u_k))| \cdot |\mathcal{L}(T_2(x)) \cap \mathcal{L}(\overline{T_1(u_2)})| \\ \cdot (|\mathcal{L}(T_2(v)) \cap \mathcal{L}(T_1(u_k))| - |\mathcal{L}(T_2(x)) \cap \mathcal{L}(T_1(u_k))|).$$

*Proof.* Consider those triplets  $X$  for which  $T_1|X$  is induced by every edge  $(u_{i-1}, u_i)$ , for  $2 \leq i \leq k$ , in  $T_1$ , and  $T_2|X$  is a subtree of  $T_2(v)$ . Let us call these triplets *relevant*. Any relevant triplet must have all three leaves from  $\mathcal{L}(T_2(v))$ , two leaves from  $\mathcal{L}(T_1(u_k))$ , and the third leaf from  $\mathcal{L}(\overline{T_1(u_2)})$ . Also note that any triplet that satisfies these three conditions must be relevant. The number of triplets that satisfy these conditions is exactly  $n_1(u_1, u_k, v)$ .

Any relevant triplet  $X$  must belong to one of the following four categories:

1. *The lca of  $X$  in  $T_2$  is not node  $v$*  : This implies that, in addition to being a relevant triplet, all three leaves of  $X$  must belong to the same subtree of  $T_2$  rooted at a child of  $v$ . The number of such triplets is  $n_2(u_1, u_k, v)$ .
2. *The lca of  $X$  in  $T_2$  is node  $v$ ,  $X$  is resolved in  $T_2$  and  $T_1|X = T_2|X$*  : A relevant triplet  $X$  satisfies this criterion if and only if there exists a child  $x \in \text{Ch}(v)$ , such that the two leaves of this triplet that belong to  $\mathcal{L}(T_1(u_k))$  in tree  $T_1$  also occur in  $\mathcal{L}(T_2(x))$ , and, the third leaf (which occurs in  $\mathcal{L}(\overline{T_1(u_2)})$  in  $T_1$ ) occurs in  $\mathcal{L}(T_2(y))$  where  $y \in \text{Ch}(v) \setminus \{x\}$ . The number of such  $X$  is equal to  $n_3(u_1, u_k, v)$ .
3. *The lca of  $X$  in  $T_2$  is node  $v$ ,  $X$  is resolved in  $T_2$ , but  $T_1|X \neq T_2|X$*  : A relevant triplet  $X$  satisfies this criterion if and only if there exists a child  $x \in \text{Ch}(v)$ , such that a pair of the leaves of  $X$  that occur in  $\mathcal{L}(T_1(u_k))$  and  $\mathcal{L}(\overline{T_1(u_2)})$  respectively in tree  $T_1$  occur in  $\mathcal{L}(T_2(x))$  in tree  $T_2$ , and, the third leaf (which occurs in  $\mathcal{L}(T_2(x))$  in  $T_1$ ) occurs in  $\mathcal{L}(T_2(y))$  where  $y \in \text{Ch}(v) \setminus \{x\}$ . The number of such  $X$  is given by  $n_4(u_1, u_k, v)$ .
4. *The lca of  $X$  in  $T_2$  is node  $v$ , and  $X$  is unresolved in  $T_2$*  : By definition, the number of relevant triplets that satisfy this criterion is exactly  $\gamma(u_1, u_k, v)$ .

We have shown that  $n_2(u_1, u_k, v)$ ,  $n_3(u_1, u_k, v)$ , and  $n_4(u_1, u_k, v)$  are exactly the number of relevant triplets belonging to categories 1, 2, and 3 respectively. The lemma follows.  $\square$

We should remark that the procedure to compute the value of  $\gamma(u_1, u_k, v)$  given in the preceding proof may seem circuitous. However, we have been unable to find a direct method with an equally good time complexity.

**Lemma 7.10.** *Given two phylogenetic trees  $T_1$  and  $T_2$  on the same  $n$  leaves, the value  $|\mathcal{R}_1(T_1, T_2)|$  can be computed in  $O(n^2)$  time.*

*Proof.* Our algorithm for computing  $|\mathcal{R}_1(T_1, T_2)|$  appears in Figure 2. The correctness of the algorithm follows from Lemma 7.7. We now analyze its complexity. For any given candidate nodes  $u, v$ , Lemma 7.9 shows how to compute  $\gamma(\cdot, \cdot, v)$  in  $O(|\text{Ch}(v)|)$  time, and consequently, by Lemma 7.8, the value  $r_1(u, v)$  can be computed in  $O(|\text{Ch}(u)| \cdot |\text{Ch}(v)|)$  time. Thus, the total time complexity of the algorithm is  $O(\sum_{u \in \mathcal{V}(T_1)} \sum_{v \in \mathcal{V}(T_2)} |\text{Ch}(u)| \cdot |\text{Ch}(v)|)$ , which is  $O(n^2)$ .  $\square$

```

procedure  $\mathcal{R}_1(T_1, T_2)$ 
1: for each internal node  $u \in \mathcal{V}(T_1) \setminus \{rt(T_1)\}$  do
2:   for each internal unresolved node  $v \in \mathcal{V}(T_2)$  do
3:     Compute  $r_1(u, v)$ .
4: return the sum of all computed  $r_1(\cdot, \cdot)$ .

```

Figure 2: Computing  $|\mathcal{R}_1(T_1, T_2)|$

## 8 An approximation algorithm for parametric quartet distance

We now consider the problem of computing the parametric quartet distance (PQD) between two unrooted trees. Our main result is an  $O(n^2)$ -time 2-approximate algorithm for PQD.

Our approach is similar to the one for computing the parametric triplet distance. Observe that Proposition 7.1 and, thus, Equation (31) hold even when the unit of distance is quartets instead of triplets. Christiansen et al. [12] show how to compute the values  $|\mathcal{S}(T_1, T_2)|$ ,  $|R(T_1)|$ ,  $|U(T_1)|$ , and  $|U(T_2)|$  within  $O(n^2)$  time. In Section 8.1 we show how to compute, in  $O(n^2)$  time, a value  $y$  such that  $|\mathcal{R}_1(T_1, T_2)| \leq y \leq 2|\mathcal{R}_1(T_1, T_2)|$ . Now, let us substitute the values of  $|R(T_1)|$ ,  $|U(T_1)|$ ,  $|U(T_2)|$  and  $|\mathcal{S}(T_1, T_2)|$  into Equation (31), and use the value of  $y$  instead of  $|\mathcal{R}_1(T_1, T_2)|$ . Assuming  $p \geq 1/2$ , it can be seen that the result is a 2-approximation to  $d^{(p)}(T_1, T_2)$ .

To summarize, we have the following result.

**Theorem 8.1.** *Given two unrooted phylogenetic trees  $T_1$  and  $T_2$  on the same  $n$  leaves, and a parameter  $p \geq 1/2$ , a value  $x$  such that  $d^{(p)}(T_1, T_2) \leq x \leq 2 \cdot d^{(p)}(T_1, T_2)$  can be computed in  $O(n^2)$  time.*

We note that the  $(2p - 1) \cdot |\mathcal{R}_1(T_1, T_2)|$  term in Equation (31) vanishes when  $p = \frac{1}{2}$ . In this case, we do not even need to compute  $|\mathcal{R}_1(T_1, T_2)|$  to get the *exact* value of  $d^{(p)}(T_1, T_2)$ .

### 8.1 Computing a 2-approximate value of $|\mathcal{R}_1(T_1, T_2)|$

For any node  $u$  in  $T$ , let  $adj(u)$  denote the set of nodes that are adjacent to  $u$ . For the purposes of describing our algorithm, it is useful to view each (undirected) edge  $\{u, v\} \in \mathcal{E}(T)$  as two directed edges  $(u, v)$  and  $(v, u)$ . Let  $\vec{\mathcal{E}}(T)$  denote the set of directed edges in tree  $T$ .

To achieve the claimed time complexity, our algorithm relies on a preprocessing step which computes and stores, for each pair of directed edges  $(u_1, v_1) \in \vec{\mathcal{E}}(T_1)$  and  $(u_2, v_2) \in \vec{\mathcal{E}}(T_2)$ , the quantity  $|\mathcal{L}(T_1(u_1, v_1)) \cap \mathcal{L}(T_2(u_2, v_2))|$ . This can be accomplished in  $O(n^2)$  by arbitrarily rooting  $T_1$  and  $T_2$  at any internal node and proceeding as in the preprocessing step for the triplet distance case (see Section 7.1).

Consider any two leaves  $a, b$  from  $\mathcal{L}(T(u, v))$  and any two leaves  $c, d$  from  $\mathcal{L}(T(v, u))$ . Then, the quartet  $\{a, b, c, d\}$  must appear resolved as  $ab|cd$  in  $T$ ; we say that the quartet tree  $ab|cd$  is *induced* by the edge  $(u, v)$ . Note that the same resolved quartet tree may be induced by multiple edges in  $T$ . Additionally, if  $x \in u_1$  and  $y \in u_2$  for some  $u_1, u_2 \in adj(u) \setminus \{v\}$  such that  $u_1 \neq u_2$ , then we say that the quartet tree  $ab|cd$  is *strictly induced* by the directed edge  $(u, v)$ .

Consider a quartet  $\{a, b, c, d\}$ . Then, the corresponding quartet tree is unresolved in  $T$  if and only if there exists exactly one node  $w$  such that the paths from  $w$  to  $a$ ,  $w$  to  $b$ ,  $w$  to  $c$ , and  $w$  to  $d$  do not share any edges. We say that quartet  $\{a, b, c, d\}$  is *associated* with node  $w$  in  $T$ . Thus, each unresolved quartet tree from  $T$  is associated with exactly one node in  $T$ .

For any directed edge  $(u, v) \in \vec{\mathcal{E}}(T_1)$  and  $w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_1)$ , let  $r_1((u, v), w)$  denote the number of quartets  $X$  such that  $T_1|X$  is strictly induced by the directed edge  $(u, v)$  in  $T_1$ , and  $X$  is associated with the node  $w$  in  $T_2$ . The quartets counted in  $r_1((u, v), w)$  must be resolved in  $T_1$  but unresolved in  $T_2$ . We have the following result.

**Lemma 8.1.** *Given  $T_1$  and  $T_2$ , we have*

$$2 \cdot |\mathcal{R}_1(T_1, T_2)| = \sum_{\substack{(u,v) \in \vec{\mathcal{E}}(T_1), \\ w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)}} r_1((u, v), w).$$

*Proof.* Let  $X = \{a, b, c, d\}$  be any quartet in  $|\mathcal{R}_1(T_1, T_2)|$ . Without loss of generality, assume that  $T_1|X = ab|cd$ , and that  $X$  is associated with node  $w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)$ . Since  $X$  appears resolved in  $T_1$ ,  $\vec{\mathcal{E}}(T_1)$  must have exactly two directed edges, say  $(u_1, v_1)$  and  $(u_2, v_2)$ , which strictly induce  $ab|cd$ . Thus,  $X$  is counted in exactly two of the  $r_1(\cdot, \cdot)$ 's, namely,  $r_1((u_1, v_1), w)$ , and  $r_1((u_2, v_2), w)$ . The lemma follows.  $\square$

Thus, we can compute  $|\mathcal{R}_1(T_1, T_2)|$  by computing all the  $O(n^2)$  possible  $r_1((u, v), w)$ 's. However, doing so seems to require at least  $\Theta(n^2 \cdot d)$  time, where  $d$  is the degree of  $T_1$ . Instead, our algorithm computes a 2-approximate value of  $|\mathcal{R}_1(T_1, T_2)|$  in  $O(n^2)$  time by relying on the next lemma.

**Lemma 8.2.** *Given  $T_1$  and  $T_2$ , let  $T'_1$  denote the rooted tree obtained from  $T_1$  by designating any internal node in  $\mathcal{V}(T_1)$  as the root. Then,*

$$|\mathcal{R}_1(T_1, T_2)| \leq \sum_{\substack{u \in \mathcal{V}(T'_1) \setminus (rt(T'_1) \cup \mathcal{L}(T'_1)), \\ w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)}} r_1((u, pa(u)), w) \leq 2 \cdot |\mathcal{R}_1(T_1, T_2)|.$$

*Proof.* First, observe that if  $u \in \mathcal{L}(T'_1)$  and  $w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)$ , then  $r_1((u, pa(u)), w) = 0$ . Therefore, we must have

$$\sum_{\substack{u \in \mathcal{V}(T'_1) \setminus (rt(T'_1) \cup \mathcal{L}(T'_1)), \\ w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)}} r_1((u, pa(u)), w) = \sum_{\substack{u \in \mathcal{V}(T'_1) \setminus rt(T'_1), \\ w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)}} r_1((u, pa(u)), w).$$

Second, observe that  $\mathcal{E}(T_1) = \mathcal{E}(T'_1)$  and, therefore, by Lemma 8.1, we must have

$$\sum_{\substack{u \in \mathcal{V}(T'_1) \setminus rt(T'_1), \\ w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)}} r_1((u, pa(u)), w) \leq 2 \cdot |\mathcal{R}_1(T_1, T_2)|.$$

This proves the second inequality in the lemma.

To complete the proof, we now prove the first inequality. Let  $X = \{a, b, c, d\}$  be any quartet in  $|\mathcal{R}_1(T_1, T_2)|$ , and, without loss of generality, assume that  $T_1|X = ab|cd$ , and that  $X$  is associated with node  $w \in V(T_2) \setminus \mathcal{L}(T_2)$ . Since  $X$  appears resolved in  $T_1$ ,  $\mathcal{E}(T_1)$  must have exactly two directed edges, say  $(u_1, v_1)$  and  $(u_2, v_2)$ , which strictly induce  $ab|cd$ . Consider the edge  $\{u_1, v_1\} \in \mathcal{E}(T_1')$ . There are two possible cases: Either  $v_1 = pa(u_1)$ , or  $u_1 = pa(v_1)$ . If  $v_1 = pa(u_1)$  then the quartet  $X$  will be counted in the value  $r_1((u_1, pa(u_1)), w)$ . Otherwise, if  $u_1 = pa(v_1)$ , then  $u_1, v_1, v_2, u_2$  must appear on a same root-to-leaf path in  $T_1'$ . Consequently, we must have  $v_2 = pa(u_2)$  and the quartet  $X$  would be counted in the value  $r_1((u_2, pa(u_2)), w)$ . Thus, we must have  $|\mathcal{R}_1(T_1, T_2)| \leq \sum_{u \in \mathcal{V}(T_1') \setminus \text{rt}(T_1')} \sum_{w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)} r_1((u, pa(u)), w)$ . The lemma follows.  $\square$

Thus, the idea for efficiently computing a 2-approximate value of  $|\mathcal{R}_1(T_1, T_2)|$  is to first root  $T_1$  arbitrarily at any internal node and then compute the value  $r_1((u, pa(u)), w)$  for each non-root node  $u \in V(T_1)$  and each  $w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_1)$ .

We now direct our attention to the problem of efficiently computing all the required values  $r_1(\cdot, \cdot)$ . Given a path  $u_1, u_2, \dots, u_k$  in  $T_1$ , where  $k \geq 2$ , let  $\gamma(u_1, u_k, w)$  denote the number of quartets  $X$  such that  $T_1|X$  is induced in  $T_1$  by every edge  $(u_{i-1}, u_i)$ ,  $2 \leq i \leq k$ , and  $X$  is associated with node  $w$  in  $T_2$ .

The following lemma is analogous to Lemma 7.8, and shows how the value  $r_1(\cdot, \cdot)$  can be computed by first computing certain  $\gamma(\cdot, \cdot, \cdot)$  values.

**Lemma 8.3.** *Let  $(u, v) \in \mathcal{E}(T_1)$ , and  $w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)$ , then,*

$$r_1((u, v), w) = \gamma(u, v, w) - \sum_{x \in \text{adj}(u) \setminus \{v\}} \gamma(x, v, w).$$

*Proof.* Let  $X = \{a, b, c, d\}$  be a quartet that is counted in  $r_1((u, v), w)$ . Without loss of generality, let  $T_1|X = ab|cd$  such that  $a, b \in \mathcal{L}(T_1(u, v))$ . It can be verified that  $X$  must satisfy the following three conditions: (i)  $X$  must be associated with node  $w$  in  $T_2$ , (ii)  $a, b \in \mathcal{L}(T_1(u, v))$  and  $c, d \in \mathcal{L}(T_1(v, u))$ , and (iii) there must not exist any  $x \in \text{adj}(u) \setminus \{v\}$  such that  $a, b \in \mathcal{L}(T_1(x, u))$ . Moreover, observe that if there exists a quartet  $X = \{a, b, c, d\}$  that satisfies these three conditions, then  $X$  will be counted in  $r_1((u, v), w)$ ; these three conditions are thus necessary and sufficient.

Now observe that  $\gamma(u, v, w)$  counts exactly all those quartets that satisfy conditions (i) and (ii), while  $\sum_{x \in \text{Ch}(u)} \gamma(pa(u), x, v)$  counts exactly all those quartets that satisfy conditions (i) and (ii), but not condition (iii). The lemma follows.  $\square$

To state our next results we need the following notation. Given phylogenetic trees  $T_1$  and  $T_2$ , consider a path  $u_1, u_2, \dots, u_k$  where  $k \geq 2$ , in tree  $T_1$ , and an internal node  $w \in \mathcal{V}(T_2)$  of degree at least 4. Let  $P = \mathcal{L}(T_1(u_1, u_2))$ ,  $Q = \mathcal{L}(T_1(u_k, u_{k-1}))$  and let  $x_1, \dots, x_{|\text{adj}(w)|}$  denote the neighbors of  $w$ . Consider the quartets that are induced by every edge  $(u_{i-1}, u_i)$ ,  $2 \leq i \leq k$ , in  $T_1$ : Let us call these quartets *relevant*. Observe that a quartet is relevant if and only if it contains exactly two leaves from  $P$  and two leaves from  $Q$ . Let

1.  $n_1(u_1, u_k, w)$  denote the number of relevant quartets  $X$  for which there exists a neighbor  $x$  of  $w$  in tree  $T_2$ , such that  $X$  is completely contained in  $T_2(x, w)$ ,

2.  $n_2(u_1, u_k, w)$  denote the number of relevant quartets  $X$  for which there exist two neighbors  $x, y$  of  $w$  in tree  $T_2$ , such that  $T_2(x, w)$  contains three leaves from  $X$  and  $T_2(y, w)$  contains the other leaf,
3.  $n_3(u_1, u_k, w)$  denote the number of relevant quartets  $X$  for which there exist two neighbors  $x, y$  of  $w$  in tree  $T_2$ , such that  $T_2(x, w)$  contains two leaves from  $X$  and  $T_2(y, w)$  contains the other two leaves, and
4.  $n_4(u_1, u_k, w)$  denote the number of relevant quartets  $X$  for which there exist three neighbors  $x, y, z$  of  $w$  in tree  $T_2$ , such that  $T_2(x, w)$  contains two leaves from  $X$ ,  $T_2(y, w)$  contains one leaf from  $X$ , and  $T_2(z, w)$  contains the remaining leaf.

Then, we must have the following.

**Lemma 8.4.**

$$\gamma(u_1, u_k, w) = \binom{|P|}{2} \cdot \binom{|Q|}{2} - n_1(u_1, u_k, w) - n_2(u_1, u_k, w) - n_3(u_1, u_k, w) - n_4(u_1, u_k, w). \quad (34)$$

*Proof.* The term  $\binom{|P|}{2} \cdot \binom{|Q|}{2}$  is the number of relevant quartets. Furthermore, each relevant quartet must occur in tree  $T_2$  in exactly one of the five configurations captured by the terms  $n_1(u_1, u_k, w)$ ,  $n_2(u_1, u_k, w)$ ,  $n_3(u_1, u_k, w)$ ,  $n_4(u_1, u_k, w)$ , and  $\gamma(u_1, u_k, w)$ . The lemma follows.  $\square$

The following four lemmas show that the values of  $n_1(u_1, u_k, w)$ ,  $n_2(u_1, u_k, w)$ ,  $n_3(u_1, u_k, w)$ , and  $n_4(u_1, u_k, w)$  can be computed in  $O(|adj(w)|)$  time. The proofs of these lemmas all follow the same approach: In each case, we show that the required value can be expressed as a sum of  $O(|adj(w)|)$  quantities, every one of which can be computed in  $O(1)$  time based on the values computed in the pre-processing step.

**Lemma 8.5.** *The value  $n_1(u_1, u_k, w)$  can be computed in  $O(|adj(w)|)$  time.*

*Proof.* We will show that

$$n_1(u_1, u_k, w) = \sum_{i=1}^{|adj(w)|} \binom{|\mathcal{L}(T_2(x_i, w)) \cap P|}{2} \cdot \binom{|\mathcal{L}(T_2(x_i, w)) \cap Q|}{2}. \quad (35)$$

The right hand side of Equation (35) counts all those quartets that are completely contained in  $\mathcal{L}(T_2(x, w))$  for some  $x \in adj(w)$  and that have two elements from  $P$  and two from  $Q$ . These are exactly the quartets that must be counted in  $n_1(u_1, u_k, w)$ .  $\square$

**Lemma 8.6.** *The value  $n_2(u_1, u_k, w)$  can be computed in  $O(|adj(w)|)$  time.*

*Proof.* We will show that

$$\begin{aligned}
n_2(u_1, u_k, w) &= \sum_{i=1}^{|adj(w)|} \binom{|\mathcal{L}(T_2(x_i, w)) \cap P|}{2} \cdot |\mathcal{L}(T_2(x_i, w)) \cap Q| \cdot |\mathcal{L}(T_2(w, x_i)) \cap Q| \\
&\quad + \sum_{i=1}^{|adj(w)|} \binom{|\mathcal{L}(T_2(x_i, w)) \cap Q|}{2} \cdot |\mathcal{L}(T_2(x_i, w)) \cap P| \cdot |\mathcal{L}(T_2(w, x_i)) \cap P|. \quad (36)
\end{aligned}$$

The quartets  $X$  counted in  $n_2(u_1, u_k, w)$  are exactly those for which there exist two neighbors  $x, y$  of  $w$  such that either (i)  $X \cap \mathcal{L}(T_2(x, w))$  contains two leaves from  $P$  and one from  $Q$ , and  $X \cap \mathcal{L}(T_2(y, w))$  contains a leaf from  $Q$  or (ii)  $X \cap \mathcal{L}(T_2(x, w))$  contains two leaves from  $Q$  and one from  $P$ , and  $X \cap \mathcal{L}(T_2(y, w))$  contains a leaf from  $P$ . The first term on the right hand side of Equation (36) is exactly the number of quartets that satisfy condition (i), and the second term on the right hand side is exactly the number of quartets satisfying condition (ii).  $\square$

**Lemma 8.7.** *The value  $n_3(u_1, u_k, w)$  can be computed in  $O(|adj(w)|)$  time.*

*Proof.* We will show that

$$\begin{aligned}
n_3(u_1, u_k, w) &= \sum_{i=1}^{|adj(w)|} \left\{ \alpha - \binom{|\mathcal{L}(T_2(x_i, w)) \cap P|}{2} \right\} \cdot \binom{|\mathcal{L}(T_2(x_i, w)) \cap Q|}{2} \\
&\quad + \frac{1}{2} \sum_{i=1}^{|adj(w)|} \{ \beta - |\mathcal{L}(T_2(x_i, w)) \cap P| \cdot |\mathcal{L}(T_2(x_i, w)) \cap Q| \} \cdot |\mathcal{L}(T_2(x_i, w)) \cap P| \quad (37)
\end{aligned}$$

Where

$$\alpha = \sum_{i=1}^{|adj(w)|} \binom{|\mathcal{L}(T_2(x_i, w)) \cap P|}{2}, \quad (38)$$

$$\beta = \sum_{i=1}^{|adj(w)|} |\mathcal{L}(T_2(x_i, w)) \cap P| \cdot |\mathcal{L}(T_2(x_i, w)) \cap Q|. \quad (39)$$

The quartets  $X$  counted in  $n_3(u_1, u_k, w)$  are exactly those quartets for which there exist two neighbors  $x, y$  of  $w$  such that either (i)  $X \cap \mathcal{L}(T_2(x, w))$  contains two leaves from  $P$ , and  $T_2(y, w)$  contains two leaves from  $Q$ , or (ii)  $X \cap \mathcal{L}(T_2(x, w))$  and  $X \cap \mathcal{L}(T_2(y, w))$  both contain one leaf each from  $P$  and  $Q$ . The first term on the right hand side of Equation (37) is exactly the number of quartets that satisfy condition (i). The sum in the second term on the right hand side counts the quartets satisfying condition (ii) exactly twice each (due to the symmetry between  $x$  and  $y$  in condition (ii)). This explains the  $\frac{1}{2}$  multiplicative factor.  $\square$

**Lemma 8.8.** *The value  $n_4(u_1, u_k, w)$  can be computed in  $O(|adj(w)|)$  time.*

**procedure**  $\text{Approx-}\mathcal{R}_1(T_1, T_2)$

- 1: Convert the unrooted tree  $T_1$  into a rooted one by rooting it at any internal node.
- 2: **for** each internal node  $u \in \mathcal{V}(T_1) \setminus \text{rt}(T_1)$  **do**
- 3:     **for** each internal unresolved node  $w \in \mathcal{V}(T_2)$  **do**
- 4:         Compute  $r_1((u, \text{pa}(u)), w)$ .
- 5: **return** the sum of all computed  $r_1(\cdot, \cdot)$ .

Figure 3: Computing a 2-approximation to  $|\mathcal{R}_1(T_1, T_2)|$

*Proof.* We will show that

$$\begin{aligned}
 n_4(u_1, u_k, w) &= \sum_{i=1}^{|adj(w)|} \binom{|\mathcal{L}(T_2(x_i, w)) \cap P|}{2} \cdot \binom{|\mathcal{L}(T_2(w, x_i)) \cap Q|}{2} \\
 &\quad + \sum_{i=1}^{|adj(w)|} \binom{|\mathcal{L}(T_2(x_i, w)) \cap Q|}{2} \cdot \binom{|\mathcal{L}(T_2(w, x_i)) \cap P|}{2} \\
 &\quad + \sum_{i=1}^{|adj(w)|} |\mathcal{L}(T_2(x_i, w)) \cap P| \cdot |\mathcal{L}(T_2(x_i, w)) \cap Q| \cdot |\mathcal{L}(T_2(w, x_i)) \cap P| \cdot |\mathcal{L}(T_2(w, x_i)) \cap Q| \\
 &\quad - 2 \cdot n_3(u_1, u_k, w).
 \end{aligned} \tag{40}$$

The quartets  $X$  counted in  $n_4(u_1, u_k, w)$  are exactly those quartets for which there exist three neighbors  $x, y, z$  of  $w$  such that either (i)  $X \cap \mathcal{L}(T_2(x, w))$  contains two leaves from  $P$ , and  $T_2(y, w)$  and  $T_2(z, w)$  each contain a leaf from  $Q$ , or (ii)  $X \cap \mathcal{L}(T_2(x, w))$  contains two leaves from  $Q$ , and  $X \cap \mathcal{L}(T_2(y, w))$  and  $X \cap \mathcal{L}(T_2(z, w))$  each contain a leaf from  $P$ , or (iii)  $X \cap \mathcal{L}(T_2(x, w))$  contains a leaf from  $P$  and a leaf from  $Q$ ,  $X \cap \mathcal{L}(T_2(y, w))$  contains a leaf from  $P$ , and  $X \cap \mathcal{L}(T_2(z, w))$  contains a leaf from  $Q$ .

The first term on the right hand side of Equation (40) counts all the quartets that satisfy condition (i), and, in addition, all the quartets that satisfy condition (i) from the proof of Lemma 8.7. Similarly, the second term on the right hand side counts the quartets that satisfy condition (ii), along with all the quartets that satisfy condition (i) from the proof of Lemma 8.7. The third term on the right hand side counts those quartets that satisfy condition (iii), and also counts, exactly twice each (again due to symmetry), those that satisfy condition (ii) from the proof of Lemma 8.7. Thus, by adding the first three terms on the right hand side of Equation (40), we obtain the value  $n_4(u_1, u_k, w) + 2 \cdot n_3(u_1, u_k, w)$ .  $\square$

**Lemma 8.9.** *Given two unrooted phylogenetic trees  $T_1$  and  $T_2$  on the same size  $n$  leaf set, a value  $y$  such that  $|\mathcal{R}_1(T_1, T_2)| \leq y \leq 2 \cdot |\mathcal{R}_1(T_1, T_2)|$  can be computed in  $O(n^2)$  time.*

*Proof.* Our algorithm to compute a 2-approximate value of  $|\mathcal{R}_1(T_1, T_2)|$  is summarized in Figure 3. Lemma 8.2 immediately implies that the algorithm computes a value between  $|\mathcal{R}_1(T_1, T_2)|$  and  $2 \cdot |\mathcal{R}_1(T_1, T_2)|$ .

We now analyze the time complexity of our algorithm. By Lemmas 8.5, 8.6, 8.7, and 8.8, the values  $n_1(u_1, u_k, w)$ ,  $n_2(u_1, u_k, w)$ ,  $n_3(u_1, u_k, w)$ , and  $n_4(u_1, u_k, w)$  can all be computed within  $O(|adj(w)|)$  time. Hence, by Lemma 8.4, the value of any  $\gamma(\cdot, \cdot, w)$  can be computed in  $O(|adj(w)|)$  time. Lemma 8.3 now implies that, for any given  $(u, v) \in \vec{\mathcal{E}}(T_1)$  and  $w \in \mathcal{V}(T_2) \setminus \mathcal{L}(T_2)$ , the value  $r_1((u, v), w)$  can be computed within  $O(|adj(u)| \cdot |adj(w)|)$  time. Thus, the total time complexity of the algorithm is  $O(\sum_{u \in \mathcal{V}(T_1)} \sum_{w \in \mathcal{V}(T_2)} |Ch(u)| \cdot |adj(w)|)$ , which is  $O(n^2)$ .  $\square$

## 9 Discussion

We have defined and analyzed distance measures for rooted and unrooted phylogenies that account for partially-resolved nodes. A number of problems remain. First, there is the question of determining whether there exists a polynomial-time algorithm for computing the median tree with respect to parametric triplet and quartet distances. We conjecture that this problem is NP-hard. Also open is the question of whether the Hausdorff distance between partially-resolved trees is NP-hard. Finally, many (if not most) applications require the comparison of trees that do not have the same set of taxa. It would be interesting to investigate whether any of our distance measures can be extended to this setting.

## References

- [1] E. N. Adams III. N-trees as nestings: Complexity, similarity, and consensus. *J. Classification*, 3(2):299–317, 1986.
- [2] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In *STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of Computing*, pages 684–693, New York, NY, USA, 2005. ACM Press.
- [3] B. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:1–13, 2001.
- [4] J. P. Barthélemy and F. R. McMorris. The median procedure for n-trees. *Journal of Classification*, 3:329–334, 1986.
- [5] J. J. Bartholdi, C. A. Tovey, and M. A. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6:157–165, 1989.
- [6] V. Berry, T. Jiang, P. E. Kearney, M. Li, and H. T. Wareham. Quartet cleaning: Improved algorithms and simulations. In *Proceedings of the 7th Annual European Symposium on Algorithms*, volume 1643 of *LNCS*, pages 313–324. Springer, 1999.
- [7] O. R. P. Bininda-Emonds, editor. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, volume 4 of *Series on Computational Biology*. Springer, 2004.



- [8] G. S. Brodal, R. Fagerberg, and C. N. S. Pedersen. Computing the quartet distance in time  $O(n \log n)$ . *Algorithmica*, 38(2):377–395, 2003.
- [9] D. Bryant. *Building trees, hunting for trees, and comparing trees: Theory and methods in phylogenetic analysis*. PhD thesis, Department of Mathematics, University of Canterbury, New Zealand, 1997.
- [10] D. Bryant. A classification of consensus methods for phylogenetics. In M. Janowitz, F.-J. Lapointe, F. McMorris, B. B. Mirkin, and F. Roberts, editors, *Bioconsensus*, volume 61 of *Discrete Mathematics and Theoretical Computer Science*, pages 163–185. American Mathematical Society, Providence, RI, 2003.
- [11] D. Bryant, J. Tsang, P. Kearney, and M. Li. Computing the quartet distance between evolutionary trees. In *SODA '00: Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 285–286, Philadelphia, PA, USA, 2000. Society for Industrial and Applied Mathematics.
- [12] C. Christiansen, T. Mailund, C. N. Pedersen, M. Randers, and M. S. Stissing. Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology*, 1(16), 2006.
- [13] J. A. Cotton, C. S. Slater, and M. Wilkinson. Discriminating supported and unsupported relationships in supertrees using triplets. *Systematic Biology*, 55(2):345–350, April 2006.
- [14] D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data*, volume 34 of *Lecture Notes in Statist.* Springer-Verlag, Berlin, 1980.
- [15] W. H. E. Day. Analysis of quartet dissimilarity measures between undirected phylogenetic trees. *Systematic Zoology*, 35(3):325–333, Sep. 1986.
- [16] P. Diaconis and R. Graham. Spearman’s footrule as a measure of disarray. *J. of the Royal Statistical Society, Series B*, 39(2):262–268, 1977.
- [17] A. C. Driskell, C. Ané, J. G. Burleigh, M. M. McMahon, B. C. O’Meara, and M. J. Sanderson. Prospects for building the tree of life from large sequence databases. *Science*, 306(5699):1172–1174, November 2004.
- [18] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Tenth International World Wide Web Conference*, pages 613–622, Hong Kong, May 2001.
- [19] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM J. Discrete Math.*, 20(3):628–648, 2006.
- [20] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top  $k$  lists. *SIAM J. Discrete Math.*, pages 134–160, 2003.

- [21] M. Farach and M. Thorup. Optimal evolutionary tree comparison by sparse dynamic programming. In *Proc. 35th Annual Symposium on Foundations of Computer Science*, pages 770–779, Piscataway, NJ, 1994. IEEE Computer Society Press.
- [22] J. Felsenstein. *Inferring Phylogenies*. Sinauer Assoc., Sunderland, Mass, 2003.
- [23] C. R. Finden and A. D. Gordon. Obtaining common pruned trees. *J. Classification*, 2(1):225–276, 1985.
- [24] S. Kannan, T. Warnow, and S. Yooseph. Computing the local consensus of trees. *SIAM J. Comput.*, 27(6):1695–1724, December 1998.
- [25] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88:577–591, 1959.
- [26] W. P. Maddison. Reconstructing character evolution on polytomous cladograms. *Cladistics*, 5:365–377, 1989.
- [27] F. R. McMorris, D. B. Meronk, and D. A. Neumann. A view of some consensus methods for trees. In J. Felsenstein, editor, *Numerical Taxonomy*, pages 122–125. Springer-Verlag, 1983.
- [28] W. Piel, M. Sanderson, M. Donoghue, and M. Walsh. Treebase. <http://www.treebase.org>. Last accessed 2 February 2007.
- [29] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
- [30] C. Semple and M. Steel. *Phylogenetics*. Oxford Lecture Series in Mathematics. Oxford University Press, Oxford, 2003.
- [31] S. Snir and S. Rao. Using max cut to enhance rooted trees consistency. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 3(4):323–333, 2006.
- [32] M. Steel and D. Penny. Distributions of tree comparison metrics — some new results. *Systematic Biology*, 42(2):126–141, 1993.
- [33] M. A. Steel. *Distributions on bicoloured evolutionary trees*. PhD thesis, Massey University, 1989.
- [34] M. Stissing, C. N. S. Pedersen, T. Mailund, G. S. Brodal, and R. Fagerberg. Computing the quartet distance between evolutionary trees of bounded degree. In D. Sankoff, L. Wang, and F. Chin, editors, *APBC*, volume 5 of *Advances in Bioinformatics and Computational Biology*, pages 101–110. Imperial College Press, 2007.
- [35] C. Stockham, L.-S. Wang, and T. Warnow. Statistically based postprocessing of phylogenetic analysis by clustering. In *ISMB*, pages 285–293, 2002.
- [36] V. V. Vazirani. *Approximation Algorithms*. Springer-Verlag, Berlin, 2001.